Towards a Comprehensive Description of the Human Retinal Transcriptome:

Identification and Characterization of Differentially Expressed Genes

Dissertation zur Erlangung des
naturwissenschaftlichen Doktorgrades
der Bayerischen Julius-Maximilians-Universität Würzburg

vorgelegt von
Heidi Schulz
aus Argentinien

Würzburg 2003

Eingereicht am 10. September 2003
Bei der Fakultät für Biologie

Mitglieder der Promotionskommission
Vorsitzender: Prof. Dr. R. Hedrich
Gutachter: Prof. Dr. Bernhard H.F. Weber
Gutachter: Prof. Dr. Ricardo Benavente

Tag des Promotionskolloquiums:
Doktorurkunde ausgehändigt am:

Erklärung gemäß §4, Absatz 3 der Promotionsordnung für die Fakultät für Biologie der Universität Würzburg:

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbständig durchgeführt und verfasst habe.
Andere Quellen als die angegebenen Hilfsmittel und Quellen wurden nicht verwendet.
Die Dissertation wurde weder in gleicher noch in ähnlicher Form in einem anderen Prüfungsverfahren vorgelegt.
Es wurde zuvor kein anderer akademischer Grad erworben.

Die vorliegende Arbeit wurde am Institut für Humangenetik der Universität Würzburg unter der Leitung von Prof. Bernhard H.F. Weber angefertigt.

Würzburg, den 10. September 2003

ACKNOWLEDGMENTS

I would like to thank Prof. Bernhard Weber for giving me the possibility to do my Ph.D. research in his group and for coaching me in the fine art of scientific investigation and writing.

My gratitude also extends to Prof. Ricardo Benavente who as a member of the Faculty of Biology of the University of Würzburg accepted to be supervisor of this thesis.

I would like to acknowledge Dr. Heidi Stöhr for introducing me to the world of gene identification and characterization.

I would also like to express my gratitude to past and present lab members of the AG Weber who helped me and contributed to a very pleasant working atmosphere. In particular I thank Jelena Stojic for the great working relationship, Vladimir Milenkovic for his patient assistance in all computer-related problems, Susanne Fröhlich and Claudia Berger for helping with the RT-PCR expression analyses, Franziska Krämer and Christine Wiedemann for the proofreading assistance, and Andrea Rivera for always being there.

I am deeply indebted to my family. You really deserve far more credit than I can ever give you! Just to mention a few things, I thank you for your generous love, support, advice, and for teaching me not with words but with your actions and life. Your integrity and warmth provided a wonderful environment in which to grow.

My most heartfelt thanks go to my husband Lucio. Without his strength, patience, and comprehension, it would not have been possible for me to finish this dissertation. Through the hardships of graduate studies, my resolve remained strong because of his love and understanding.

My final recognition and thanks go to God for His everlasting love, guidance, the health He gave me these years, and for being my strength.

Table of Contents

I Summary

The human retina is a multilayered neuroectodermal tissue specialized in the transformation of light energy into electric impulses which can be transmitted to the brain where they are perceived as vision. Since the retina is easily accessible and functional aspects are directly recordable, the study of this tissue has been at the forefront of neuroscience research for over a century. Studies have revealed that the distinct functions of the retina require a large degree of differentiation which is achieved by the coordinated function of approximately 55 different cell types. The highly structured anatomy and the functional differentiation of the retina is a result of its distinctive transcriptome and proteome.

Due to the complexity of the retina it has been difficult to estimate the number of genes actively transcribed in this tissue. Great efforts in the elucidation of retinal disease genes have led to the identification of 139 retina disease loci with 90 of the corresponding genes cloned thus far[1]. In contrast to the success in the hereditary disorders, efforts to identify the genetic factors conferring manifestations known as age-related macular degeneration (AMD) have revealed sparse results. AMD is a retinal disease affecting a significant percentage of the older population. This disorder is likely due to exogenic as well as genetic factors.

To further our understanding of retinal physiology and facilitate the identification of genes underlying retinal degenerations, particularly AMD, our efforts concentrated on the systematic analysis of the retinal transcriptome. Since approximately half of all retinal degeneration-associated genes identified to date are preferentially expressed in retina, it is plausible that the investigation of gene expression profiles and the identification of retina-expressed transcripts could be an important starting point for characterizing candidate genes for the retinal diseases. The expressed sequence tags approach included the assessment of all retinal expressed sequence tags (EST) clusters indexed in the UniGene database and of 1080 single-pass ESTs derived from an in-house generated human retina suppression subtracted hybridization (SSH) cDNA library. In total, 6603 EST clusters were evaluated during this thesis and detailed *in-silico* analysis was performed on 750 EST clusters. The expression of the genes was evaluated using reverse transcriptase-polymerase chain reaction (RT-PCR), followed by confirmation using quantitative reverse transcriptase-polymerase chain reaction (qRT-PCR), as well as conventional and virtual Northern blot analysis. The expression profiling of 337 selected EST clusters led to the identification of 111 transcripts, of which 60 are specific or abundant to the retina, 3 are expressed at high levels in the retinal pigment epithelium (RPE), and 48 are expressed in brain as well as in retina.

The EST approach used to select candidate transcripts allowed us to assess the effectiveness of the two available resources, the UniGene database and the retinal SSH (retSSH) cDNA library. From the results obtained, it is evident that the generation of suppression subtracted libraries to identify cell-specific transcripts constitutes the most straight-forward and efficient strategy. In addition to the high

[1] http://www.sph.uth.tmc.edu/Retnet

percentage of candidate genes that are identified from an SSH cDNA library, it has the added benefit that genes expressed at low levels can be identified. Furthermore, comparison of our retina-enriched gene set with previously published studies demonstrated only limited overlap of the identified genes further confirming the valuable source of retinal genes from our retinal SSH cDNA library.

The effort of our and other groups has resulted in the establishment of the full-length coding sequence of 55 of the 111 genes uniquely or preferentially expressed in the retina. Using various methods such as bioinformatical analysis, EST assembly, cDNA library screening, and rapid amplification of cDNA ends (RACE) a number of genes were cloned in the scope of this thesis including C1orf32, C4orf11, C7orf9, C12orf7, C14orf29, DAPL1, and GRM7.

Bioinformatic analyses and cDNA library screening were used to isolate the full-length cDNA sequence and determine the genomic organization of C7orf9, also identified as RFRP. This 1190 bp retina-specific transcript from chromosome 7p15.3 encodes a precursor protein for at least two small neuropeptides, referred to as RFRP-1 and RFRP-3. Since C7orf9 is localized in the critical region for dominant cystoid macular dystrophy (CYMD) its role in the pathology was investigated. Southern blot analysis and sequencing of samples from two affected individuals of the original pedigree used to localize the disease gene excluded the gene from involvement in this disease.

Multiple isoforms of the C12orf7 gene were assembled from a number of clones identified from library screenings, PCR amplifications, and RACE experiments. The gene variants, transcribed from chromosome 12q13.13, have been found to be expressed exclusively in retina. Because of the multiple alternative splicing of the gene, we can only speculate about the nature of the protein it encodes. The longest transcript, which includes all six exons plus the last intervening sequence, encodes a 471 aa protein which contains a nuclear localization signal and five ankyrin repeats. The existence of many isoforms is also observed in mouse suggesting that they may have a relevant role in cellular physiology.

Five novel splice variants of the glutamate metabotropic receptor 7 (GRM7) resulting from the use of alternative 3'-end exons were identified and characterized. One of the novel variants, GRM7_v3, encodes a 924 aa protein and is therefore the longest putative GRM7 protein reported to date. Even though they are not retina-specific, the isoforms are preferentially expressed in the nervous system. Although the functional properties of the specific carboxyl-termini are still unclear, it is known that axon targeting of GRM7_v1 is mediated by the last 60 aa of the protein. Hence the novel isoforms may direct the protein to specific subcellular localizations.

The C1orf32 gene, preferentially expressed in retina, is organized in 10 exons and is transcribed from chromosome 1q24.1. Bioinformatic analyses of the 639 aa putative protein not only identified the mouse and rat orthologous genes but also the LISCH7 gene as a potential member of the same family. Since the LISCH7 protein has been shown to function as a low density lipoprotein receptor, the C1orf32 protein may be involved in retinal lipid homeostasis. Disturbances in lipid metabolism have

been proposed as one of the pathways involved in AMD etiology. Thus, the role of C1orf32 in this complex disease should be investigated.

Expression analyses of the death-associated protein-like 1 (DAPL1) gene revealed that it is expressed in both the retina and the RPE at high levels. The 552 bp transcript encodes a 107 aa putative protein and is transcribed from chromosome 2q24.1. *In-silico* analyses identified an additional 12 related proteins from various species which share high similarity constituting a novel protein family. The similarity to the death-associated-protein (DAP) is particularly interesting since this protein has been found to be indispensable for programmed cell death. Therefore, DAPL1 is an excellent candidate for retinal disease as apoptosis is generally the ultimate cause in retinal degeneration.

The retina-specific C4orf11 and C14orf29 genes localized on chromosome 4q21.22 and 14q22.1, respectively, are both transcribed in more than one isoform. The encoded proteins do not contain any known domains but because of their retina-specific expression they may be important for proper retinal physiology.

As part of the long-term goals of the project, several of the cloned genes are being genotyped to construct single nucleotide polymorphism (SNP) maps. Projects to investigate haplotype frequencies of candidate genes in large cohorts of controls and AMD patients are ongoing. Thus, by establishing a collection of 111 genes expressed exclusively or preferentially in the retina, the present work has laid the foundation for future research in retinal diseases.

II Zusammenfassung

Die menschliche Retina ist ein mehrschichtiges neuroektodermales Gewebe, das auf die Umwandlung von Lichtenergie in elektrische Impulse spezialisiert ist. Diese Impulse werden zum Gehirn weitergeleitet, wo sie als Bilder gesehen werden. Da die Retina experimentell leicht zugänglich ist und funktionelle Aspekte direkt untersucht werden können, nehmen Experimente an diesem Gewebe in der neurologischen Forschung seit mehr als einem Jahrhundert einen Spitzenplatz ein. Es wurde gezeigt, dass die unterschiedlichen Funktionen der Retina durch eine enorme Differenzierung in mehr als 55 Zelltypen, die koordiniert miteinander in Kontakt treten, ermöglicht wird. Die hohe strukturelle und funktionelle Komplexität der Retina wiederum ist das Resultat ihres Transkriptoms und ihres Proteoms.

Die Komplexität der Retina erschwert es, die Zahl aktiv transkribierter Gene in diesem Gewebe zu schätzen. Durch die vielfältigen Bemühungen, Gene zu identifizieren, die mit retinalen Erkrankungen in Zusammenhang stehen, konnten bisher 139 Genloci mit retinalen Krankheiten in Verbindung gebracht werden und in 90 Fällen wurden die jeweiligen Gene bereits kloniert[2]. Während auf dem Gebiet der hereditären Erkrankungen damit bereits große Erfolge erzielt wurden, sind die Resultate, was komplexe Erkrankungen wie die altersabhängige Makuladegeneration (AMD) betrifft, noch spärlich. AMD ist eine retinale Erkrankung, die einen signifikanten Prozentsatz der älteren Bevölkerung betrifft. Es wird angenommen, dass sowohl exogene als auf genetische Faktoren als Auslöser beteiligt sind.

Um die Physiologie der Retina besser zu verstehen und die Identifikation von Genen, die in retinale Erkrankungen involviert sind, zu ermöglichen, wurde in dieser Arbeit ein Schwerpunkt auf die systematische Analyse des retinalen Transkriptoms gelegt. Etwa die Hälfte aller Gene, die bei retinalen Erkrankungen eine Rolle spielen und bis heute identifiziert wurden, werden überwiegend in der Retina exprimiert. Das Erstellen von Genexpressionsprofilen und die Identifikation von retina-spezifischen Transkripten ist daher der erste wichtige Schritt für die Charakterisierung von Kandidatengenen für retinale Erkrankungen. Zu diesem Zweck wurde eine „Expressed Sequence Tags“ (ESTs) Analyse durchgeführt, in der alle Retina „Clusters“ der UniGene Datenbank sowie 1080 einzelne ESTs einer laboreigenen, suppressiv-subtraktiv hybridisierten (SSH) cDNA Bank der menschlichen Retina bewertet. Insgesamt wurden 6603 EST „Clusters“ während dieser Arbeit untersucht und für 750 eine detaillierte *in-silico* Analyse angeschlossen. Die Expression der Gene wurde mittels reverser Transkriptase-Polymerase-Kettenreaktion (RT-PCR) untersucht und mit quantitativer RT-PCR (qRT-PCR) bestätigt. Außerdem wurden konventionelle und virtuelle Northern Blot Hybridisierungen zur Aufklärung der Expressionsprofile eingesetzt. Die Expressionsprofile von 337 EST „clustern“ führte zur Identifikation von 111 Transkripten, von denen 60 abundant oder spezifisch in der Retina exprimiert werden, drei eine hohe Expression im retinalen Pigmentepithel (RPE) zeigen und 48 sowohl in der Retina als auch im Gehirn vorkommen.

[2] http://www.sph.uth.tmc.edu/Retnet

Die Bewertung der ESTs zur Auswahl von Kandidatengenen erlaubte einen direkten Vergleich der beiden genutzten Datenressoucen, der UniGene Datenbank und der retinalen SSH (retSSH) cDNA Bank. Die Resultate zeigten, dass die Generierung einer SSH Bank zur Identifikation zellspezifischer Transkripte die effektivere Methode darstellt. Durch die Nutzung dieser cDNA Bank konnte nicht nur ein Großteil der Kandidatengene, sondern auch Gene mit einer niedrigen Expression identifiziert werden. Außerdem zeigt die cDNA Bank eine Anreicherung an Retina-Transkripten. Somit stellt die retSSH cDNA Bank eine wertvolle Quelle zu Identifizierung neuer retinaler Gene dar.

Durch diese und andere Arbeiten konnte die vollständige kodierende Sequenz von 55 der 111 retinaspezifischen oder –abundanten Gene ermittelt werden. Mittels verschiedener Methoden wie bioinformatische Analysen, „EST assembly“, cDNA Bank „screening“ und „RACE-Experimenten“ konnten im Lauf dieser Arbeit mehrere Gene, darunter C1orf32, C4orf11, C7orf9, C12orf7, C14orf29, DAPL1 und GRM7, kloniert werden.

Mit bioinformatischen Analysen und cDNA Bank „screening“ wurde die Volllänge cDNA Sequenz und die genomische Organisation von C7orf9 (alias RFRP), ermittelt. Das 1190 bp retinaspezifische Transkript auf Chromosom 7p15.3 kodiert ein Vorläuferprotein für mindestens zwei kleine Neuropeptide, RFRP-1 und RFRP-3. Da C7orf9 in einer kritischen Region für die dominante zystoide Makuladystrophie liegt, wurde eine mögliche Rolle für die Pathologie der Krankheit untersucht. Mittels Southern Blot Hybridisierungen und der Sequenzierung von C7orf9 zweier betroffener Patienten des gleichen Stammbaums, der für die Lokalisierung des Krankheitsgens verwendet worden war, konnte eine Beteiligung des Genes am Krankheitsbild ausgeschlossen werden.

Mehrere Isoformen des C12orf7 Gens konnten mit Hilfe verschiedener Klone aus cDNA Bank „screeining“, PCR Amplifikationen und „RACE-Experimenten“ identifiziert werden. Die von Chromosom 12q13.13 transkribierten Genvarianten zeigen eine retinaspezifische Expression. Aufgrund der vielfältigen Spleißvarianten des Gens kann über die Eigenschaften des kodierten Proteins nur spekuliert werden. Das längste Transkript, das alle sechs Exone und die letzte Intron-Sequenz enthält, kodiert ein Protein mit 471 AS, das ein Kernlokalisierungssignal und fünf Ankyrin „Domains“ aufweist. Die Existenz mehrere Isoformen wurde auch in der Maus nachgewiesen, was auf eine funktionelle Relevanz in der Physiologie der Zelle hinweist.

Fünf neue Spleißvarianten des Glutamat metabotropen Rezeptors 7 (GRM7) mit alternativen Exonen am 3´Ende wurden identifiziert und charakterisiert. Eine der neuen Varianten, GRM7_v3 kodiert ein 924 AS-langer Protein und stellt damit das mutmaßlich längste GRM7 Protein dar. Die Isoformen werden nicht retinaspezifisch, sondern verstärkt in neuronalen Gewebe exprimiert. Obwohl die funktionellen Eigenschaften der spezifischen Carboxyl-Enden noch nicht geklärt sind, ist bekannt, dass „axon-targeting“ von GRM7_v1 von den letzten 60 AS des Proteins vermittelt wird. Die neuen Isoformen könnten daher eine Rolle im subzellulären Transport des Proteins spielen.

Das C1orf32 Gen, das vorwiegend in der Retina exprimiert wird, besitzt zehn Exone und liegt auf Chromosom 1q24.1. Mit bioinformatische Analysen konnten nicht nur die orthologen Gene des putativen Proteins mit 639 AS in Maus und Ratte, sondern auch das LISCH7 Gen als potentielles Mitglied der Genfamilie identifiziert werden. Für LISCH7 ist eine Funktion als „low density lipoprotein receptor“ bekannt, das C1orf32 Protein ist damit möglicherweise in den retinalen Fettstoffwechsel involviert. Störungen im Fettstoffwechsel sind als Ursache für AMD vorgeschlagen worden, die Rolle von C1orf32 in dieser komplexen Krankheit sollte deshalb untersucht werden.

Expressionsanalysen des „death-associated protein-like 1“ (DAPL1) Gens zeigten eine hohe Expression sowohl in Retina als auch im retinales Pigmentepithel. Das 552 bp Transkript kodiert ein Protein mit 107 AS und liegt auf Chromosom 2q24.1. *In-silico* Analysen identifizierten 12 weitere verwandte Proteine verschiedener Spezies, die eine hohe Ähnlichkeit aufweisen und eine neue Proteinfamilie darstellen. Vor allem die Ähnlichkeit mit dem „death-associated“ Protein, das für den programmierten Zelltod essentiell ist, ist von besonderem Interesse. Apoptose ist meist der ultimative Grund der retinalen Degeneration und somit stellt DAPL1 ein ausgezeichnetes Kandidatengen für Retinopathien dar.

C4orf11, lokalisiert auf Chromosom 4q21.22 und C14orf2, lokalisiert auf Chromosom 14q22.1, sind retinaspezifische Genen und werden in mehren Isoformen transkribiert. Die kodierten Proteine enthalten keine bekannten Domänen, wegen ihrer retinaspezifischen Transkription kann allerdings eine wichtige Rolle für die Funktion der Retina nicht ausgeschlossen werden.

Als eines der langfristigen Ziele dieses Projekts wurden mehrere der geklonten Gene genotypisiert um „single nucleotide polymorphism“ (SNP) Karten zu erstellen. Die derzeitigen Projekte untersuchen die Haplotypfrequenzen der Kandidatengene in großen Kohorten von AMD Patienten und nicht betroffenen Kontrollpersonen. Die Kollektion von 111 retinaspezifischen oder –abundanten Genen, die in dieser Arbeit identifiziert und charakterisiert wurden, hat damit eine wichtige Grundlage für die weitere Erforschung retinaler Erkrankungen gelegt.

III Introduction

1. The human retina

Vision plays a key role in our interaction with the environment and although it is one of five senses, approximately 38% of the neuronal input to the brain comes from visual information (Alward 2003). The light signals which are captured by the eye not only play a role in vision, but also in circadian rhythms, regulation of body color, and detection of seasonal changes.

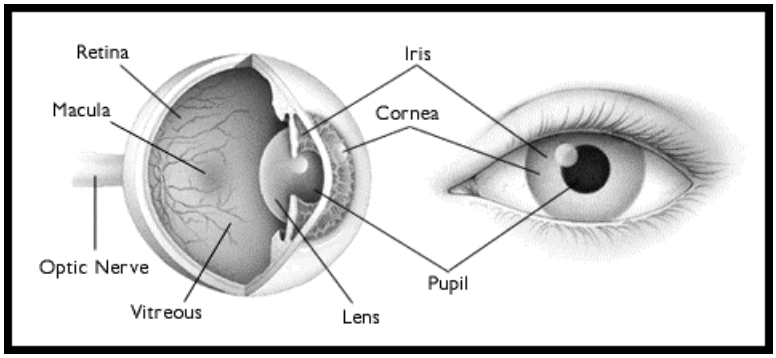


Fig. 1 Anatomy of the eye
Light enters the eye and passes through the cornea, lens and vitreous before reaching the retina. The highest visual acuity is achieved in the central area of the retina, called macula. In the retina light impulses are converted into electrical impulses which are conveyed to the brain via the optic nerve.
(Drawing reproduced from http://www.eyesight.org/All_About_MD/all_about_ md.html)

The retina is a highly ordered tissue of neuroectodermal origin which covers two-thirds of the inner eye (Fig. 1). Its function is the conversion of light into electrical signals which are relayed via the bipolar cells to the ganglion cells. The visual information is then carried through the axons of the ganglion cells, which form the optic nerve, to the higher visual centers in the brain.

The highly specialized function of photoreception, initial visual processing, and integration are accomplished by the complex and delicate organization of 55 cell types (Masland 2001) organized in ten distinct laminae constituting the retina (Fig. 1). The major cell types include the photoreceptors, the glial Müller cells, and the horizontal, bipolar, amacrine, and ganglion neurons (Fig. 2). Mature photoreceptors are highly polarized neurons that contain four distinct compartments: the outer segment, the inner segment, a connecting cilium, and a cell body containing the nucleus. The outer segments of both photoreceptor types, the rods and the cones, consist of stacks of coin-like discs whose membranes contain the visual pigment. In the case of rods it is rhodopsin, whereas iopsin is found in the cones. The inner segment contains most of the metabolic machinery and cellular components; the metabolites are exchanged between the inner and outer segment through the cilium (Gordon and Bazan 1997).

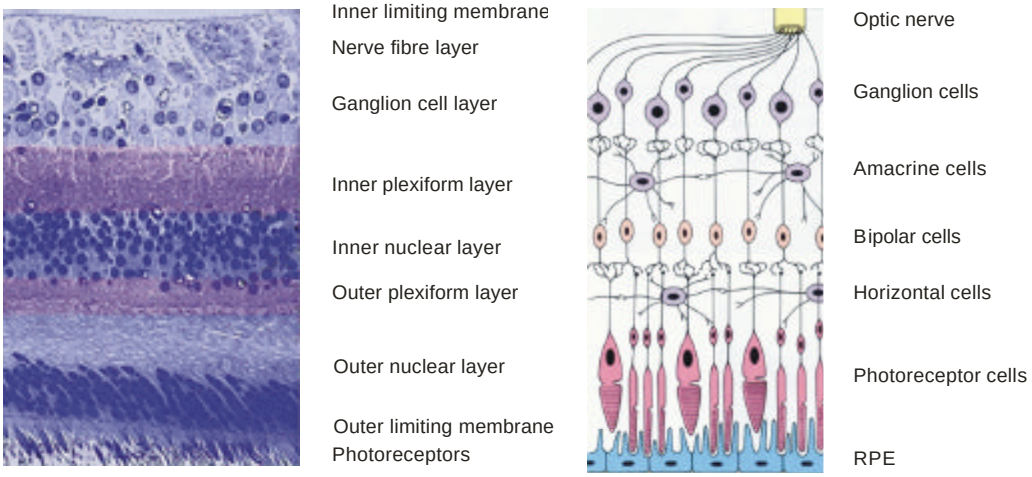


Fig. 2 Organization of the human retina
The histological organization of the retina in distinct layers is shown on the left picture. The RPE layer, found beneath the photoreceptors is not depicted. The schema on the right shows the approximate form of the cells and their localization. Picture reproduced from http://webvision.med.utah.edu/sretina.html. The schema is reproduced from Spalton et al. 1993.

The distribution of the photoreceptors within the retina is not uniform. An accumulation of cones is found in the macula lutea (Fig. 1), which has a diameter of approximately 5.5 - 6.0 mm and is characterized by a yellow pigmentation due to its high lutein and zeaxanthin content (Spalton et al. 1993). The fovea centralis, a depression in the macula is located in the center of the visual axis and has a diameter of 1.5 mm. With 150 thousand cones per mm^2 the fovea contains the highest concentration of cones. The six million cones are responsible for high-resolution visual acuity and color vision (Curcio et al. 1990). The specification of wavelength is achieved by the existence of three types of cones: long-, medium-, and short-wave (Bowmaker and Dartnall 1980). In contrast to the cones, the highest concentration of rods is found in the periphery. The 120 million rods function in conditions of low illumination and enable peripheral and night vision.

The outer segments of the photoreceptors are in close proximity to the single-cell layer of the retinal pigment epithelium (RPE) and although apical prolongations of the RPE engulf the photoreceptors there is no anatomical bond between them. An important anatomical aspect of the RPE are the tight junctions between the monolayer cells which form the zonular occludens. These tight junctions constitute the ‘outer blood-retinal barrier’. Major functions of the RPE include absorption of stray light, transportation of metabolites and nutrients and phagocytosis of the discs continually shed by the photoreceptor cells (Spalton et al. 1993). Whereas the rod discs are replaced at a rate of 10 - 15% daily, it takes the cone discs nine to twelve months to be totally substituted (Gregory et al. 1991). The shedded discs are not always completely metabolized by the RPE but may be deposited during lifetime as cellular waste (lipofuscin) in the RPE and thus can influence the correct function of the cell (Krott and Heimann 1996). Between the RPE and the underlying choriocapillaris lies Bruch’s membrane which consists of five layers. The choroid blood vessels of the choriocapillaris are essential for the blood supply of the RPE and the photoreceptors.

2. Hereditary retinal degenerations

The highly evolved architecture of the retina makes it particularly susceptible to a great number of genetic defects. Retinal degenerations are the most common form of hereditary eye disorders and affect approximately 1 in 3000 persons (Inglehearn 1998). Different parameters used to classify them include the age of onset (congenital, early or late in life), the mode of inheritance (autosomal dominant or recessive, X-linked), the underlying genetic defects (monogenic vs. multigenic), the clinical symptoms, or the affected anatomical region (peripheral versus central retina degenerations).

The most widely used classification is based on the affected region. In this classification, retinal degenerations are categorized as 'generalized' if the degeneration eventually affects the function of the whole retina or as 'regionally-restricted' if the damage remains localized (Kellner 1997). In the two categories the initial lesions may be in the center or in the periphery of the retina. Disorders which primarily affect the rod system result in loss of night vision and may progress to affect peripheral vision (Table 1). They are usually referred to as retinitis pigmentosa (RP). Diseases which affect the central cone-rich area can be due to dysfunction of the cones, alteration of the RPE, or the choriocapillaris and will lead to a progressive loss of central vision (Table 1). If all three types of cone photoreceptors die, the affected person develops achromatopsia, a condition in which vision is mediated exclusively by rod photoreceptors.

Table 1 Categorization of the principal retinal degenerations

Retinal degenerations with peripheral onset

Generalized		Regional	
Name	**Inheritance**	**Name**	**Inheritance**
Retinitis pigmentosa	AD, AR, X, M, D	Autosomal dominant vitreoretinochoroidopathy	AD
Usher syndrome	AR		
Leber congenital amaurosis	AR		
Congenital stationary night blindness	AD, AR, X		

Retinal degenerations with central onset

Generalized		Regional	
Name	**Inheritance**	**Name**	**Inheritance**
Sorsby's fundus dystrophy	AD	Stargardt disease	AD, AR
Cone dystrophy	AD, AR, X	Best dystrophy	AD
Cone-rod dystrophy	AD, AR, X	X-chromosomal congenital retinoschisis	X
		North Carolina macular dystrophy	AD
		Age-related macula degeneration	C

AD: autosomal dominant; AR: autosomal recessive; X: X-linked; M: mitochondrial; C: complex inheritance; D: di-allelic
Table was compiled according to information published in Kellner (1997)

2.1. The age-related macular degeneration (AMD)

The majority of the retinal degenerations generally manifest in early adulthood while the most common form of macula degeneration has a late onset and has therefore been denominated age-related macular degeneration (AMD). This disorder, first described in 1855 (Donders 1855), has a multifactorial etiology and is caused by exogenous as well as genetic factors. It usually develops in the sixth or seventh decade of life (Wood 2000). It is the most important cause of new cases of blindness in patients over 55 years of age and considering the demographic change of Western societies, an estimated 16 million persons will be affected by the year 2030 (Evans and Wormald 1996). Its etiology remains elusive, but interplay between environmental and genetic factors is thought to be critical to the development of the disease. The incidence varies somewhat between different ethnic groups (Evans 2001) but in general 20% of the persons aged 65 to 74 are affected and the percentage increases to 35% in the group of 75 to 84 years olds (Schick et al. 2001).

Characteristic features of AMD include accumulation of membranous debris on both sides of the RPE basement membrane and growth of abnormal blood vessels into the subretinal macula (Zarbin 1998). The progression of AMD can be divided into an early and a late stage. The early stage is characterized by the presence of large small yellowish deposits (drusen) and pigmentary abnormalities in the macular area but only low visual impairment (Yates and Moore 2000). The late stage is subdivided in two types, the dry and the wet forms. Only 15% of AMD patients develop the wet form (Seddon 2001) which is responsible for the majority of cases of severe loss of central vision.

Current evidence suggests that AMD is probably triggered by environmental factors in genetically susceptible subjects. Reported risk factors include smoking (Seddon et al. 1996, Christen et al. 1996, Klein et al. 1998), increased exposure to sunlight (Cruickshanks et al. 1993), low plasma concentrations of antioxidant vitamins and zinc (Vanden Langenberg et al. 1998), as well as gender (Klein et al. 1997), ethnic origin (Klein et al. 1995, Schachat et al. 1995, Friedman at al. 1999), and iris color (Mitchell et al. 1998). Several studies have attempted to dissect the genetic component of the susceptibility to AMD but until now no single gene has been found to be a major risk factor (reviewed in Stöhr 2003). Instead AMD probably results from the contribution of several genes exhibiting low or moderate effects.

3. Molecular genetics of human retinal disease

The efforts to identify the genes responsible for hereditary diseases began in the early 1900s with the identification of Mendelian-like inheritance of ‘inborn errors of metabolism’. Interestingly, the first gene to be mapped to a specific chromosome in any species was the one for colorblindness (Wilson 1911). Although the majority of the retinal degenerations are inherited following classical Mendelian inheritance rules, the identification of the molecular basis of the disorders is complicated by the existence of genetic heterogeneity (different genes causing the same disease), allelic heterogeneity (different mutations in the same gene causing either the same or different diseases), and clinical heterogeneity (same mutations in a single gene cause different phenotypes).

As a result of the advancements in the genetic field it has been possible to identify 139 loci which harbour genes related to retinal disease and clone 90 of these genes. From the 139 retinal degeneration loci, 45 are inherited in an autosomal dominant mode, 68 in an autosomal recessive mode, 21 are X-linked, and 5 are encoded in the mitochondrial genome (Table 2).

Table 2 Number of loci and genes involved in various retinal degenerations

Disease	No. of loci	No. of cloned genes
Bardet-Biedl syndrome, autosomal recessive	7	5
Cone or cone-rod dystrophy, autosomal dominant	7	4
Cone or cone-rod dystrophy, autosomal recessive	2	0
Cone or cone-rod dystrophy, X-linked	2	0
Congenital stationary night blindness, autosomal dominant	1	1
Congenital stationary night blindness, autosomal recessive	2	2
Congenital stationary night blindness, X-linked	2	2
Leber congenital amaurosis, autosomal recessive	7	4
Macular degeneration, autosomal dominant	10	4
Macular degeneration, autosomal recessive	1	1
Ocular-retinal developmental disease, autosomal dominant	1	0
Optic atrophy, autosomal dominant	2	1
Optic atrophy, X-linked	1	0
Retinitis pigmentosa, autosomal dominant	12	11
Retinitis pigmentosa, autosomal recessive	15	10
Retinitis pigmentosa, X-linked	5	2
Syndromic or systemic retinopathy, autosomal dominant	3	2
Syndromic or systemic retinopathy, autosomal recessive	12	9
Syndromic or systemic retinopathy, X-linked	2	1
Usher syndrome, autosomal recessive	11	7
Other retinopathy, autosomal dominant	9	4
Other retinopathy, autosomal recessive	11	8
Other retinopathy, mitochondrial	5	5
Other retinopathy, X-linked	9	7
Total	**139**	**90**

Table was extracted from the RetNet database[3]

The most genetically heterogeneous disorder is RP. More than twenty causative genes have been cloned and an additional thirteen genes have been mapped (Wang et al. 2001). Classic examples of clinical heterogeneity are a deletion in the RDS gene that can lead to retinitis pigmentosa, pattern dystrophy or fundus flavimaculatus (Weleber et al. 1993) or mutations in rhodopsin which may cause dominant RP (Dryja et al. 1990), recessive RP (Rosenfeld et al. 1992), or dominant stationary night blindness (Dryja et al. 1996). Various methods have been applied to identify the genes involved in retinal diseases. A brief description of each will be presented in the following section.

4. Gene identification approaches for monogenic disorders

Unravelling the genetic basis of human disease is probably 'the central challenge' of modern human genetics. Since this enterprise can be difficult and time-consuming (Fig. 3) several strategies for the identification of disease genes have been developed. They can be classified in two main groups. The

[3] http://www.sph.uth.tmc.edu/Retnet/sum-dis.htm

first includes those approaches that identify the disease gene based on information regarding its chromosomal location (positional cloning) whereas the second group includes all position-independent techniques (sequence homology or functional complementation).

Fig. 3 Public campaign about gene identification
Advertisement campaign of the Hospital for Sick Children in Toronto, Canada. The first interactive wall in the world was designed to publicize the work that is done within the Hospital. Two viewfinders (the 'microscopes') were located on the sidewalk in front of the wall. Passers-by could use them to scan the wall for the hidden pins. Each pin highlights one of 10 genes discovered at HSC. For example, beneath one pin it read '1989: Cystic Fibrosis'.

The first method used to identify disease genes was based on the study of cytogenetic abnormalities present in patients. But since such aberrations are rare, linkage analysis has been the primary method for mapping Mendelian disorders. Linkage analysis tests for co-segregation of a gene marker and disease phenotype within a family are used to determine if the marker and the disease locus are physically linked. Commonly used markers include restriction fragment length polymorphisms (Botstein et al. 1980), simple sequence repeats (Weber and May 1989, Litt and Luty 1989), or single nucleotide polymorphisms (Kuklin et al. 1997). Typically, positional data delimits a chromosomal region of approximately 1-5 cM that usually contains between 20 and 200 genes. The choice then lies between sequencing large number of genes or setting priorities by combining positional data with available expression and phenotype data. In the ophthalmologic field a number of genes have been identified using this approach. Classic examples of successful positional cloning include the identification of the VMD2 (Marquardt et al. 1998), NDP (Berger et al. 1992), and EFEMP1 (Stone et al. 1999) genes which are involved in Best disease, Norrie disease, and Doyne honeycomb retinal degeneration, respectively.

Whereas linkage analysis depends on the co-segregation of a gene (locus) and a phenotype through a pedigree, allelic association analysis, or linkage disequilibrium mapping, relies on measuring deviation from the random occurrence of alleles in unrelated patients or nuclear families versus control individuals (Collins and Morton 1998). Since the introduction of single nucleotide polymorphisms (SNPs) to research, this gene-identification strategy has gained popularity, particularly for the investigation of complex diseases. The retinal degeneration research field has also benefited from this method since the molecular basis of Usher syndrome type 3 was discovered by haplotype and linkage-disequilibrium analysis (Joensuu et al. 2001).

Despite the recent successes, a number of disease genes that remain to be discovered might not be as accessible. A proportion of genes may contribute only minimally to a disease phenotype and thus may not be identified with the discussed approaches. With the advent of the genome project there has

been a shift from the 'phenotype-based' towards the 'gene-based' methods. These methods begin with the identification, cloning, and characterization of a gene. Only after there is enough information about the function and/or expression of the gene, efforts are then directed towards elucidating what human phenotype could result from an allelic variation of the particular gene. The sequences of candidate genes are obtained from intra- or inter-species sequence comparisons, identification of transcribed sequences from specially constructed libraries, study of model animals, or *in-silico* predictions.

The optimal selection of candidate genes is usually achieved by choosing genes that are part of a physiological pathway known to play a role in the trait or genes that may be key to the normal function of the particular tissue due to their particular expression profile. The use of expression criteria to select candidate genes has been particularly popularized in the last years and for example has been applied to identify prostate (Vasmatzis et al. 1998), heart (Mégy et al. 2002), and brain (Qiu et al. 2002) genes with relevant roles in the corresponding tissue. The expressed sequence tag (EST) approach has also been successfully used to identify apoptosis-related genes of the cardiovascular system (Rezvani et al. 2000) and novel human signalling proteins (Schultz et al. 2000).

In retinal research, the use of a gene-based approach led, for example, to the association of the ABCA4 gene to Stargardt disease (Allikments et al. 1996). The ABCA4 gene, member of the ATP-binding cassette (ABC) superfamily, was considered a candidate gene since it localized to the chromosomal region delineated to contain the Stargardt disease gene and was exclusively expressed in retina. Sequencing of the gene confirmed the relationship between the gene and Stargardt disease. Another example of the use of this strategy was the characterization of the RPE65 gene which is expressed specifically in the RPE. Suspecting that it would be a promising candidate gene for retinal dystrophies, Morimura et al. (1998) investigated the RPE65 gene in a number of retinal degeneration patients and found that it is indeed involved in RP and Leber congenital amaurosis. Sequence comparison is also becoming more feasible to identify novel candidate genes. For example, the characterization of the retina-specific retinitis pigmentosa 1-like 1 (RP1L1) gene is the result of the realization that there was a sequence highly homologous to the RP1 gene, which is a frequent cause of autosomal dominant RP (Bowne et al. 2003).

5. Gene identification approaches for complex disorders

The approaches described in the previous section have led to great advances in the cloning of genes responsible for monogenic diseases. In contrast, many of the fundamental questions relating to the genetic etiology of human diseases which are of greater public health concern remain unanswered (Altmüller et al. 2001 and Pritchard and Cox 2002). A considerable proportion of these diseases appear to aggregate within families but do not segregate following a strictly Mendelian mode of inheritance (Fig. 4). Typical examples include cardiovascular diseases, nutritional disorders (obesity, diabetes), auto-immune diseases (multiple sclerosis), psychiatric disorders (schizophrenia, bipolar disorders, depression, dementia), degenerative disorders (Parkinson or Alzheimer disease), and

retinal degenerations (AMD). The study and discovery of the molecular bases of these disorders is cumbersome mainly due to four reasons. First, they typically vary in symptom severity and age of onset, thus, it is difficult to define a phenotype and select the best population to study. Second, diverse etiological mechanisms may lead to closely overlapping phenotype. Another factor is that they are more likely to be caused by several genes, each with a small overall contribution and relative risk. For example, in major cancers and heart disease highly penetrant mutations are rare and appear to account for less than 5% of the cases (Winkelmann et al. 2000). Finally, there is a significant environmental contribution (Tabor et al. 2002) with up to 80-90% of the individual relative risk due to external factors (Armstrong 1975 and Kato et al. 1973). Hence, this group of diseases are defined as 'polygenic', 'multifactorial', 'complex', or 'multigenic' (Weeks and Lathrop 1995).

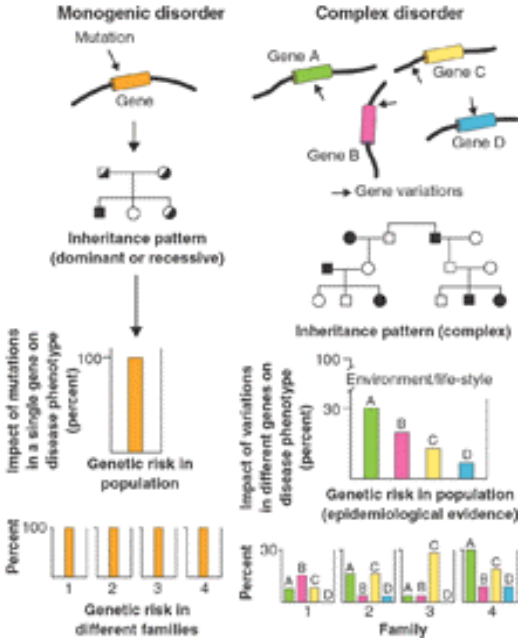


Fig. 4 Inheritance models of monogenic and complex disorders
In monogenic diseases, mutations in a single gene are sufficient to produce the clinical phenotype and to cause the disease. The impact of the gene on genetic risk for the disease is the same in all families. In complex disorders with multiple causes, variations in a number of genes encoding different proteins result in a genetic predisposition to a clinical phenotype. Pedigrees reveal no classic Mendelian inheritance pattern, and gene mutations are often neither sufficient nor necessary to explain the disease pheno-type.

Environment and life-style are major contributors to the pathogenesis of complex diseases. In a given population, epidemiological studies expose the relative impact of individual genes on the disease phenotype. However, between families the impact of these same genes might be totally different. In one family, a rare gene C (Family 3) might have a large impact on genetic predisposition to a disease. However, because of its rarity in the general population, the overall population effect of this gene would be small.

(Reprinted with permission from Peltonen and McKusik (2001) Science 291:1224-1229. Copyright 2001. American Association for the Advancement of Science)

Although linkage mapping is very successful when applied to rare monogenic diseases, few common diseases have been mapped to statistical significance using exclusively this approach. Thus, the trend is to tackle the genetic causes of common diseases by identifying susceptibility regions using linkage analysis, refining the localization with linkage disequilibrium, and studying candidate genes on a large scale. This has led to significant successes in the identification of the susceptibility genes of such complex traits as Crohn's disease (Hugot et al. 2001 and Ogura et al. 2001), Alzheimer's disease (Corder et al. 1993 and Raffai et al. 2001), and type 2 diabetes (Horikawa et al. 2000). Even though the investigation of complex diseases is slowly producing successful results, a review of 101 reports of complex disease studies could not find any common clear strategy for success (Altmüller et al. 2001).

The search for the genetic components underlying AMD susceptibility has been tackled by using two strategies, namely genome-wide searches for susceptibility loci and investigation of candidate genes. In a genome-wide study, locus 1q25-q31 was identified as a possible susceptibility region in a large family affected by AMD (Klein et al. 1998). This locus was independently confirmed in a large sib-pair analysis which also identified three additional probable AMD loci (at 1q31, 9p13, 10q26, and 17q25)

(Weeks et al. 2001). In a study of 70 families with affected members, loci 3p13, 4q32, 9q33, and 10q26 were found to reveal probably association with AMD (Majewski et al. 2003). The 10q26 locus was also found in the study by Weeks (2000). Recently, three new loci, at 3q26, 12q23, and 16p12 have been proposed by Schick et al. (2003).

In the second approach, several candidate genes have been investigated in order to determine whether allelic variants or mutations in the genes are genetically associated with AMD. The tendency has been to select genes that may be of pathophysiological relevance for AMD (e.g. APOE, OPTC, CTS3, ACE) or associated with other macula dystrophies (e.g. ABCA4, TIMP3, VMD2, ELOVL4). Numerous studies in APOE and ABCA4 (Allikmets et al. 1997, Souied et al. 1998, Klaver et al. 1998, Souied et al. 2000, Simonelli et al. 2001, and Schmidt et al. 2002) indicate that these genes may be involved in the pathomechanisms of AMD, but larger population studies are required to confirm this. For the other genes, differences in allelic variance frequency were found but, until now, no proof of the role of one of these genes in AMD could be established (reviewed in Stöhr 2003).

The current situation points out the need to identify additional candidate genes for testing their role in AMD susceptibility. Since the 'low-hanging-fruit' genes have already been cloned, new technologies and approaches will be necessary to identify novel retinal genes and begin to establish the retinal transcriptome (Risch 2000). For these efforts, the availability of genome sequences and advances in data management and retrieval are of invaluable importance.

6. Contributions of the genomic era to gene identification efforts

The modern era of genetics started just 137 years ago with the publication of Gregor Mendel's work (Mendel 1866). In just over a century scientists have unravelled many mysteries, from the organization and workings of the nucleus to the mechanisms by which genetics may be used to understand and treat diseases. In this process, a myriad of methods and tools have been developed to aid the investigations. One of the most outstanding developments, achieved in 1977, was the development by W. Gilbert and F. Sanger of the technology to determine the exact nucleotide sequence of DNA. This was followed by the description of the method to copy DNA using polymerase chain reaction (PCR) in 1983 by K. Mullis (Fig. 5).

In light of all the advances, the relentless drive to decipher genes and entire genomes culminated in a series of genome sequencing projects, the most notorious of which was the Human Genome Project begun in 1990. The first milestone of the genome projects was achieved in 1995 with the publication of the first complete bacterial genome, the *H. influenzae* sequence (Fleischmann et al. 1995) (Fig. 5). This was followed by the unveiling of a series of genomes from model organisms, including *S. cervisiae* which was the first eukaryotic genome to be sequenced (Fig. 5). Only four years later the draft sequence of the human genome was published (Lander et al. 2001 and Venter et al. 2001). Another milestone which would play an essential role in the identification of genes was the proposal to sequence the ends of cloned cDNA sequences (Adams et al. 1991), a method which was thereon

identified as expressed sequence tags (EST). The use of ESTs has since then been used to clone novel genes, analyze gene expression, and identify nucleotide polymorphisms.

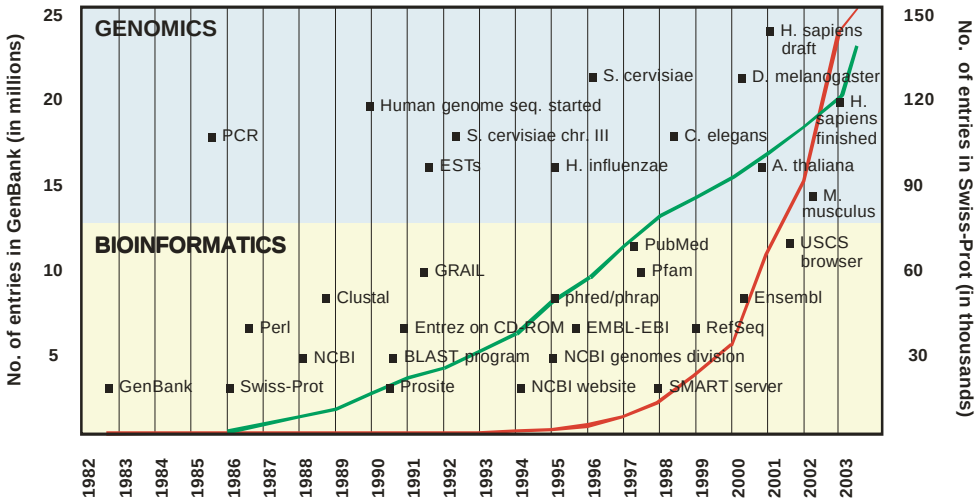


Fig. 5 Milestones in genomics and bioinformatics
The timeline shows significant discoveries and developments in the bioinformatic and genomic fields. An indication of the repercussion of the achievements is the exponential accumulation of DNA sequence information in GenBank[4] (red line) and protein sequences in Swiss-Prot[5] (green line). Except otherwise noted, each of the achievements is plotted in the year in which it was discovered other culminated. (Modified from Wolfe and Li 2003 with copyright permission from Nature Publishing Group[6])

The genome projects have changed our concept of biology and the methods used for the investigations, with a shift from laboratory work to the analysis of information. The management and analysis of the boundless amount of information being generated required the blending of specialties to originate a new field: bioinformatics. As the name suggests, it encompasses the use of tools and techniques from molecular biology, computing, statistics, mathematics, and data analysis algorithms. One of the main applications of bioinformatics is data mining, which is defined as the nontrivial extraction of implicit, previously unknown and potentially useful information, or the search for relationships and global patterns that exist in databases (Frawley et al. 1992). Since the introduction of sequence databases (e.g. GenBank[7] and Swiss-Prot[8]) and the BLAST algorithm (Altschul et al. 1990), the availability of bioinformatic tools has played a key role in the annotation of the genome and discovery of novel genes (Fig. 5). And so, just 13 years after the beginning of the genome sequencing efforts, 1540 from the 30,000 genes predicted to be encoded in the human genome have been linked to a disease[9].

The realization that the human genome encodes only approximately 30,000 genes (Lander et al. 2001 and Venter et al. 2001) and contains a greater than thought quantity of non-coding sequence that are transcribed (Mattick and Gagen 2001 and Shabalina et al. 2001) has raised many questions. The

[4] http://www.ncbi.nlm.nih.gov/Genbank/index.html
[5] http://www.expasy.org/sprot/relnotes/relstat.html
[6] http://www.nature.com/
[7] http://www.ncbi.nlm.nih.gov/Genbank/index.html
[8] http://www.expasy.org/sprot/relnotes/relstat.html
[9] Number of genes obtained by a LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/) query of human entries with the parameters 'disease_known AND has_seq"

dilemma of how human complexity can be achieved with such a limited number of genes has been partially resolved by the suggestion that there may be many more mechanisms to regulate gene expression than previously thought. These include alternative promoter usage, post-transcriptional modifications, and alternative splicing (Su et al. 2002, Modrek and Lee 2002). With this in mind, the compilation of the transcriptome, which is defined as the collection of all transcribed elements of the genome in a particular setting at a certain time point, becomes of utmost importance. It includes not only all expressed mRNAs, but also non-coding RNAs and transcribed pseudogenes. The importance and benefits of defining the transcriptome of a cell are gaining preponderance even though for many years it was postulated that the ultimate goal of the post-genomic era would be to characterize the proteome or collection of proteins found in a cell. The fact is that mRNAs not only define the proteome but are important regulatory agents (Eddy 2001, Szymanski and Barciszewski 2002). Because of its versatile and important functions, RNA was proclaimed the Breakthrough of the Year 2002 (Couzin 2002).

All cells contain a 'minimal' transcriptome which is shared between all tissues and a set of specific transcripts which are unique to each tissue. It is exactly the identification of the specific transcripts that is of interest for disease-gene identification projects because these transcripts may correlate with the fundamental physiological processes unique to a tissue. Advances in genomic research have made it possible to identify many of the ubiquitously or abundantly expressed genes. Thus at present, the greatest hurdle is the identification of the set of genes expressed at low levels in each cell. An estimated 70% of all protein-coding human genes are found only in one transcript per cell posing a major challenge for the comprehensive characterization of the cellular transcriptome of a given tissue (Kuznetsov 2002).

Efforts to decipher particular transcriptomes began more than a decade ago and a number of technologies suited for large-scale identification of the set of genes expressed in a cell-type have been developed. However, it has not been possible to compile a single complete transcriptome until now. The challenge was clearly stated in the subtitle of the 2001 Human Proteome Project Meeting which read: 'Genes Were Easy'. A real-life proof of this is the fact that even seven years after the completion of the genome from *S. cerevisiae*, the true size of its transcriptome is still not clear (Cliften et al. 2001 and Kumar et al. 2002). New genes are still being discovered and the existence of reported genes is being questioned.

The human retinal transcriptome is no exception. Until the end of the 1990s, significant efforts had been made to identify genes involved in retinal degenerations using the phenotype-based approach. On the contrary, a very limited collection of genes was available for gene-based approaches since no systematic analysis to characterize the retina transcriptome had been undertaken. In this context, the start of a project that would systematically investigate EST data from the retina and experimentally evaluate the expression of these transcripts in a set of tissues, in order to identify retina genes and particularly those preferentially expressed in retina, was an urgent necessity. Undoubtedly, the characterization of the retinal transcriptome will facilitate the understanding of the molecular basis of retinal function, thus providing the foundation for system biology (Kitano 2002).

7. Goal of the thesis

The human retina is known to be susceptible to a great number of genetic defects that lead to a wide range of retinal disease phenotypes. At the start of this doctoral thesis, only 58 genes from 118 identified retinal disease loci had been cloned and fewer advances had been made in dissecting the molecular basis of the multifactorial age-related macular degeneration (AMD).

Therefore, the principal goal of this project was the generation of a comprehensive catalogue of retinal transcripts that would aid the identification of genes underlying retinal degeneration and would further the understanding of retinal physiology. A systematic search for genes preferentially expressed in the retina was achieved by the use of two complementary approaches, namely the assessment of retinal expressed sequence tags (EST) clusters indexed in the UniGene database and of ESTs derived from an in-house generated human retina suppression subtracted hybridization cDNA library. After bioinformatic characterization, the expression profiles of selected EST clusters were determined by RT-PCR, conventional and virtual Northern blot, and quantitative real-time PCR. Using this approach, a number of transcripts expressed abundantly or exclusively in the human retina could be identified. The efforts were then directed towards the isolation and characterization of novel genes.

The availability of these novel retinal genes paves the way for a further goal, namely the study of the involvement of these genes in various retina degenerations with special emphasis on AMD. In particular, the generation of single nucleotide polymorphism (SNP) maps has been initiated for selected genes and will be utilized in case-control association studies for the genetically complex AMD.

IV Materials and Methods

1. Bioinformatic tools

1.1. Databases

1.1.1. Sequence databases

1.1.1.1. GenBank

GenBank®[10] is the collection of all publicly available DNA sequences. The entries are identified with an accession number and can be searched and retrieved using the Entrez platform[11]. In compliance with international scientific practice, all the sequences cloned in this project have been submitted to GenBank using the BankIt tool[12] and have received the corresponding accession numbers.

1.1.1.2. Databases of expressed sequence tags (EST)

The dbEST database (Boguski et al. 1993) is a division of GenBank® that contains EST sequence data from humans and other species along with information like tissue of origin and type of library from which a clone was derived. This database has been growing exponentially, e.g. the 4 million EST sequences deposited in dbEST at the beginning of this project have quadrupled to more than 17 million sequences of which 5 million are of human origin. The cDNA clones from which the ESTs are derived receive a name (e.g. yt72f03) and the orientation of the ESTs is indicated by the ending ‘r1’ (5’-end) or ‘s1’ (3’-end) (e.g. yt72f03.r1). In addition, a unique GenBank accession number (acc.no.) is assigned to each EST (e.g. R93826).

1.1.1.3. UniGene

UniGene is a gene-index where EST sequences sharing an identical 50 base pair overlap in the 3’-untranslated region (UTR) are grouped in a common EST cluster (Schuler et al. 1996). Each human EST cluster is identified by the letters Hs. (*Homo sapiens*) followed by an identification number (e.g. Hs.102453). Hence, each UniGene cluster contains sequences that theoretically represent a unique gene as well as related information such as the tissue in which the gene is expressed. To date 4,407,974 human ESTs are included in 108,094 Hs. UniGene clusters. Some clusters contain more than 1000 ESTs, while 31,188 consist of only one EST (singletons). The UniGene database can be queried with key word(s), GenBank accession number, and cluster identification numbers. Alternatively, it is possible to search for chromosome-specific EST clusters.

[10] http://www.ncbi.nlm.nih.gov/Genbank/index.html
[11] http://www.ncbi.nlm.nih.gov/Entrez/
[12] http://www.ncbi.nlm.nih.gov/BankIt/

1.1.1.4. TIGR

The Institute for Genome Research has constructed a strict gene index database called TIGR[13]. ESTs with a minimum 40 base pair overlap showing at least 95% sequence identity and a maximum unmatched overhang of 30 base pairs are clustered in Tentative Human Consensus sequences (THCs). Each THC receives a unique identifier (e.g. THC361432). Individual THCs can be found either by homology searches using a query sequence or by searching for their identifier. Throughout the study, this database was used as a complement of the UniGene database. The advantage of the TIGR entries is that they provide the consensus sequence of the assembled sequences plus annotations about the sequences and their position relative to the other sequences of the consensus cluster.

1.1.2. Gene information databases

Information about genes and sequences investigated in this doctoral thesis was acquired principally from three sources: the GeneCards[14], the LocusLink[15], and the SOURCE[16] collections. The GeneCards database (Rebhan et al. 1997) offers concise and well-organized information about the functions of human genes. The data is extracted from more than 30 sources dealing with human genes. The LocusLink database, which is part of the National Center for Biotechnology Information (NCBI) package, provides a single query interface to curated sequences. It also presents information on official nomenclature, aliases, sequence accession numbers, phenotypes, MIM numbers, UniGene clusters, homology, and mapping location. An advantage of this collection is the assignment of a specific LocusLink ID to each gene which can be used for reference or cross-searches in other databases. For most purposes, the two databases described above were fully sufficient to gather information about a gene. But when it was necessary to gain a fast impression about the expression of a gene, the SOURCE unification tool was used. This database dynamically collects and compiles data from many scientific databases. It is able to provide a good approximation of the *in-silico* expression of a gene because of its unique CloneReports feature which generates an *in-silico* expression profile from data available in dbEST.

1.1.3. Databases of genes involved in disease

For information about the characteristics of a disease and the gene involved in the pathology the Online Mendelian Inheritance in Man™ (OMIM) database[17] was consulted. All human genes and genetic disorders are catalogued in this collection and brief descriptions about findings as well as links to literature references, sequence records, maps, and related databases are provided. For specific information about retinal diseases and related genes the Retinal Information Network[18] (RetNet™)

[13] http://tigrblast.tigr.org/tgi/
[14] http://bioinformatics.weizmann.ac.il/cards/
[15] http://www.ncbi.nlm.nih.gov/LocusLink/index.html
[16] http://source.stanford.edu/cgi-bin/sourceSearch
[17] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
[18] http://www.sph.uth.tmc.edu/Retnet/

database was used. It provides specific information about retinal diseases and genes involved in retinal pathologies.

1.1.4. Other databases

1.1.4.1. PubMed

Literature searches and retrieval of articles found in MEDLINE and some additional journals were conducted at PubMed[19].

1.1.4.2. HUGO Gene Nomenclature (HGCN) Committee

The symbols and names for all genes assembled in this project were obtained from the HGCN committee and are stored in the Human Gene Nomenclature Database[20].

1.1.4.3. HuGEIndex Gene Specific Expression database

The HuGEIndex Gene Specific Expression database[21] was queried to compile a list of genes which might be ubiquitiously expressed or be involved in housekeeping functions. A list of 1169 genes present in colon, kidney, liver, lung, muscle, and oesophagus was retrieved and analyzed. Because many of the GenBank identifiers for the genes were outdated, the LocusID of all entries which were identifiable were retrieved and used in all future comparisons. The final housekeeping gene list was therefore reduced to 1065 entries, henceforth referred to as the housekeeping list.

1.2. Sequence analysis tools

The first step in the analysis of novel sequences usually involved the screening for repetitive elements using the Repeat Masker[22] tool (Smit and Green, unpublished data). In order to identify related or partially identical transcripts, the sequence was compared to sequences compiled in various databases (e.g. non-redundant, EST, SwissProt, etc.) using the Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990) at the platform offered by NCBI[23].

For the simultaneous analysis of more than one sequence or the compilation of information about the genomic locus of a gene, the UCSC Genome Browser[24] (Kent et al. 2002) was the method of choice. This browser, also commonly known as GoldenPath, has been available since the end of 2001 and provides a rapid and reliable display of any requested portion of the human, mouse, rat, *C. elegans*, *C. griggsae*, and SARS genomes. It supports text and sequence based searches (denominated BLAT) and provides quick and precise access to any region of interest. Compared to BLAST, it may miss

[19] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
[20] http://www.gene.ucl.ac.uk/nomenclature/
[21] http://www.hugeindex.org/
[22] http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker
[23] http://www.ncbi.nlm.nih.gov/BLAST/
[24] http://genome.ucsc.edu/

more divergent or short sequence alignments but it is able to find similar sequences in a fraction of the time needed by BLAST. Secondary links from individual entries within annotation tracks lead to sequence details and supplementary off-site databases.

A similar tool which shows the results of running many DNA analysis programs on a DNA sequence is offered by the UK HGMP Resource Center. The NIX tool[25] runs and combines the results of a series of programs such as GRAIL, Fex, Hexon, MZEF, Genemark, Genefinder, FGene, BLAST (against many databases), Polyah, RepeatMasker, and tRNAscan. An advantage over other programs is the fact that very long sequences (> 100 kb) can be investigated.

Computation of pairwise or multiple alignments of DNA or protein sequences was done by submission of the sequences to the Sequence Analysis Server[26] developed at the Department of Computer Science at Michigan Tech. The representation and visualization of sequence alignments was aided by BOXSHADE[27] (version 3.21) since it transforms multiple alignments to make them printer-friendly and adds features like shading.

Multiple alignment of phylogenetically related sequences and generation of a phylip output was done by ClustalW comparison at the European Bioinformatic Institute website[28]. Phyolgenetical trees were generated by inserting the phylip output in the Phylodendron software[29] developed by D.G. Gilbert.

The open reading frame (ORF) of mRNA transcripts was identified and translated using either the Translate tool[30] or the ORF Finder[31] which vary somewhat in their output.

In-silico protein analysis was principally done using the tools assembled at the ExPASy proteomics server (Gasteiger et al. 2003) of the Swiss Institute of Bioinformatics (SIB). This server groups dozens of programs and links to facilitate protein characterization. Just to name a few, the areas covered in this collection include translation tools, similarity, pattern, and profile searches, post-translational modification predictions, and primary structure analysis.

1.2.1. Splice site discrimination score

The probability that a donor or acceptor splice site will be used in the process of pre-mRNA processing can be inferred by analysis of its flanking sequence and comparison to the results computed for more than 700 other splice sites (Penotti 1991). Based on the study by Berg and von Hippel (1987) which established a discrimination score for each of the positions surrounding a splice site, it is possible to calculate a splice score by adding the individual scores at each nucleotide position. A score of 0 for the 5'-donor site denotes that the sequence is identical to the consensus; a

[25] http://www.hgmp.mrc.ac.uk/Registered/Webapp/nix/
[26] http://genome.cs.mtu.edu/
[27] http://www.ch.embnet.org/software/BOX_form.html
[28] http://www.ebi.ac.uk/clustalw/index.html
[29] http://iubio.bio.indiana.edu/treeapp/treeprint-form.html
[30] http://www.expasy.org/tools/dna.html
[31] http://www.ncbi.nlm.nih.gov/gorf/gorf.html

score of 30.1 indicates that there is no similarity at all with the consensus sequence. In reality, from the 764 investigated splice sites, all but one had scores between 0 and 11 (Penotti 1991). For the 3'-acceptor site the theoretical values range between 0 and 42.5, but again all but five of 764 splice sites had a theoretical below 14 (Penotti 1991). To facilitate the calculation of the scores a spreadsheet, compiled by Dr. Christian Sauer (2001), was used to automatically calculate the splice-site scores.

2. RNA-related methods

2.1. Sources

2.1.1. Tissue

Retina and retinal pigment epithelium were isolated from donor eyes obtained by the University Eye Clinic, Wuerzburg. The eye donations were in accordance with regulations and guidelines of the local ethics committee. The eyes were dissected immediately after receipt to isolate the retina and the RPE and the dissected tissues were snap-frozen by immersion in liquid nitrogen and stored at -80°C.

2.1.2. RNA

RNA from human adrenal gland, bone marrow, cerebellum, brain, colon, distal colon, fetal brain, fetal liver, heart, kidney, lung, liver, placenta prostate, salivary gland, skeletal muscle, small intestine, spinal cord, spleen, stomach, testis, thymus, thyroid, trachea, uterus, occipital cortex, basal ganglia, and bladder were acquired commercially from various companies including Ambion (Austin, USA), Research Genetics (Huntsville, USA), and Clontech (BD Biosciences Clontech, Heidelberg, Germany). Retina and RPE RNAs were isolated from the tissue donor eyes. All RNAs were stored at -80°C.

2.2. RNA isolation

Total RNA was isolated from frozen tissue using the RNeasy Mini® Kit (Qiagen, Hilden, Germany) following the instructions of the manufacturer. Briefly, the frozen tissue was homogenized in the provided highly denaturing guanidine isothiocyanate (GITC)-containing buffer, which immediately inactivates RNases to ensure isolation of intact RNA. Ethanol was then added to provide appropriate binding conditions, and the sample was loaded to an RNeasy Mini column where the total RNA bound to the membrane and contaminants were washed away. The RNA was then eluted with 30 µl of H_20 and stored at -80°C.

2.3. Elimination of contaminating genomic DNA

Small amounts of genomic DNA that are usually isolated together with the RNA were eliminated by treatment with the DNA-free^{TM} removal system (Ambion, Austin, USA). A 0.1 vol of 10x DNase I buffer (100 mM Tris-Cl - pH 7.5, 25 mM $MgCl_2$, 1 mM $CaCl_2$) was added to the RNA sample together with 2 U of DNase I and the mix was incubated at 37°C for 30 min. The DNase I was inactivated by

addition of 0.1 vol of the provided slurry to the sample. The RNA-containing supernatant was separated and stored for future applications at -20°C.

2.4. Northern blot analysis

2.4.1. Filter preparation

Total RNA (10 µg) was electrophoresed in a 1.2% (w/v) agarose/formaldehyde gel prepared by dissolving 12 g agarose in 87 ml DEPC water to which 10 ml of 10x MOPS (0.2 M MOPS, 50 mM NaOAc, 10 mM Na_2EDTA - pH 7.0) and 3 ml formaldehyde (37%) were added. Previous to loading the RNA was mixed with at least 1 vol of RNA loading buffer (1x MOPS, 50% formamide, 18.5% formaldehyde, 0.04% bromophenol blue, 10 µg ethidium bromide) and denatured for 10 min at 65°C. The size of the transcripts was determined by comparison to the RNA Standard marker (Promega, Mannheim, Germany). After running the gel for 2-3 hrs at 55V in 1x MOPS a picture with a reference ruler was taken and the gel was rinsed for 10 min in 20x SSC (3 M NaCl, 0.3 M NaCitrate; pH 7.0). The quality of the RNA was assessed by evaluation of the 28S (6.3 kb) and 18S (2.3 kb) bands. The RNA was then transferred to a Nylon Hybond N^+ membrane (Amersham Biosciences, Freiburg, Germany) using 60 mbar the vacuum apparatus VacuGene^{TM} (Amersham Biosciences, Freiburg, Germany).

2.4.2. Probe labeling

The probe used to hybridize the filter was prepared by random-prime labeling of the DNA with a^{32}P-dCTP. In short, 20-50 ng of the DNA fragment were denatured in a final volume of 34 µl for 3 min at 100°C. The labeling was achieved by addition of 10 µl of 5x OLB buffer (250 nM Tris-HCl, pH 8.0; 25 nM $MgCl_2$, 50 mM ß-mercaptoethanol, 96 µM each dATP, dGTP, dTTP, 1 M Hepes - pH 6.6; 50 U (A_{260}) pd$(N)_6$), 20 mg BSA, 4 U DNA Polymerase I, large (*Klenow*) fragment (Invitrogen, Karlsruhe, Germany), and 3 µl a^{32}P-dCTP (3000 Ci/mmol) to the sample and incubation overnight at room temperature. Residual a^{32}P-dCTP was removed from the probe using a G25 Sephadex column (Amersham Biosciences, Freiburg, Germany). Just before adding the probe to the membrane, it was denatured for 5 min at 100°C.

2.4.3. Membrane hybridization and washing

Prior to hybridization the membrane was pre-hybridized for 30 min in 25 ml Church-buffer (0.5 M Na_3PO4 - pH 7.2, 1 mM Na_2EDTA - pH 8.0, 7% SDS). After this step 10 ml of the buffer were removed and the denatured radio-labeled probe was added. The hybridization took place overnight in a rotating oven at 65°C. Unhybridized probe was removed by washing the membrane for 20 min with 50 ml of decreasing concentrations of the SSPE/0.1% SDS buffer (20x SSPE: 3 M NaCl, 200 mM NaH_2PO_4, 20 mM Na_2EDTA). The stringency of the washing is increased as the SSPE concentration is decreased. Therefore, the first rinse was done with 2x SSPE and depending on the amount of unbound probe it was stepwise reduced to 0.1x SSPE. The membrane was then sealed in a thin plastic bag and

exposed to X-ray Retina film (Fotochemische Werke GmbH, Berlin, Germany) film for 12 hrs to 5 days at -80°C or exposed for a week to a Storage Phosphor Screen (Amersham Biosciences, Freiburg, Germany).

2.5. First-strand cDNA synthesis

First-strand cDNA synthesis was carried out by mixing 10 pmol of 3'-RACE AP primer (Table 35, Appendix) with 1 µg of total RNA in a final volume of 12 µl. After a 70°C, 10 min incubation 4 µl of 5x First-Strand Buffer (250 mM Tris-HCl - pH 8.3, 375 mM KCl, 15 mM $MgCl_2$), 2 µl 0.1 M DTT, 1 µl dNTP mix (10 mM each of dATP, dGTP, dCTP and dTTP), and 200 µl of SuperScript II™ (Invitrogen, Karlsruhe, Germany) were added. After incubation at 42°C for 50 min the enzyme was inactivated by heating to 70°C for 15 min.

3. DNA-related methods

3.1. Isolation of plasmid DNA

Plasmid DNA was isolated from bacterial overnight cultures in LB (Luria-Bertani) medium using the NucleoSpin®Plasmid kit (Macherey-Nagel, Düren, Germany) which is based on the SDS/alkaline lysis method presented in Birnboim and Doly (1979). Pure plasmid DNA was obtained by following the instructions of the manufacturer.

3.2. Polymerase chain reaction (PCR)

DNA fragments were amplified by a three-step PCR. In the first step, the template DNA was heated to 94°C for 4 min in the first cycle and 30 sec in the following cycles. The second step lasted 30 sec and consisted in the annealing of the forward and reverse oligonucleotide primers to the single-stranded DNA at the specific annealing temperature of the primers (T_a = {2 x (A+T) + 4 x (C+G)} – 2°C). In the elongation step, the temperature was raised to the optimal temperature for the polymerase being used and held for a minimum of 30 sec. The extension time was prolonged for longer amplicons so that an extra minute was added per extra kb to be amplified. After 33 cycles a final extension of 5 min was carried out at the extension temperature.

All PCRs were done in a final volume of 25 µl using a master mixture (prepared in house) containing 1x buffer (50 mM KCl, 10 mM Tris-HCl - pH 8.3, 0.01% gelatine) with variable concentrations of MgCl2 (range 1.0 – 1.5 mM), 400 nM forward and reverse primers (Metabion, Martinsried, Germany; MWG Biotech, Ebersberg, Gemany; Invitrogen, Karlsruhe, Germany), 100 µM of each dNTP (Peqlab, Erlangen, Germany), and 1 U of *Taq* polymerase, plus template DNA or cDNA. For certain PCRs formamide was also added to a final concentration of 4%. For difficult PCRs or those where long products needed to be amplified the Platinum® *Taq* DNA Polymerase High Fidelity (Invitrogen, Karlsruhe, Germany), the KOD HiFi DNA Polymerase (Merck Biosciences GmbH, Schwalbach, Germany), or the Elongase® Enzyme Mix (Invitrogen, Karlsruhe, Germany) were used.

Oligonucleotide primers for non-quantitative PCR were designed using Primer3 software[32] or the OLIGO version 2.0 tool (Rychlik and Rhoads 1989). The 18-22 bp long primers usually contained a guanine or cytosine at the 3'-end. Only oligonucleotides with a melting temperature higher than 52°C, absence of hairpins at the 3'-end, and less than five inter-molecular base pair bridge were accepted. The lyophilized oligonucleotides synthesized by MWG Biotech (Ebersberg, Gemany) and Metabion (Planegg-Martinsried, Germany) were dissolved in ddH_2O to a final concentration of 100 pmol/µl and stored at -80°C. The working concentration of 10 pmol/µl was stored at -20°C.

3.3. Relative quantitative real-time PCR (qRT-PCR)

3.3.1. Primer design

Primers for qRT-PCR were designed with the MGB Eclipse™ Design software[33] (Epoch Biosciences, Bothell, USA). If possible, the primers were selected so that a fragment of 75 to 250 bp would be amplified from the last two exons of the gene, with one of the primers aligning partially to both exons. The primer-dimer formation feasibility was investigated with the NetPrimer software[34] (PREMIER Biosoft International, Palo Alto, USA). Primers were ordered from MWG Biotech (Ebersberg, Gemany), Metabion (Planegg-Martinsried, Germany) or Invitrogen (Karlsruhe, Germany).

3.3.2. Optimization of qRT-PCR reactions

The optimization of PCR conditions was carried out using the gradient function of the iCycler (Bio-Rad, Munich, Germany) or the DNA Opticon® 2 (Biozym Diagnostik GmbH, Hess. Oldendorf, Germany) systems. In general the first reaction was done using a $MgCl_2$ concentration of 3 mM and annealing temperatures of 57, 60 and 63°C. The variables that were changed in order to obtain more efficient amplifications with the least possible amount of primer dimers included the annealing temperature, the $MgCl_2$ concentration, the elimination of the extension step at 72°C, and the incorporation of a fluorescence measuring step at the highest temperature before the specific product started to denature. All products were loaded to ensure amplification of a single discrete band of the desired size. The conditions used for each primer combination are summarized in the Table 35 (Appendix).

3.3.3. Determination of amplification efficiency

To quantify the results obtained by qRT-PCR, calibration curves to determine the linearity of each reaction were generated. The curves were obtained by fivefold (panel O) or fourfold (panels H and T) serial dilutions of retina cDNA. For all primer pairs, each dilution plus a H_2O control was amplified in triplicate with each primer pair. To calculate the amplification efficiency of each primer pair each curve and its melting profile were carefully analyzed before accepting it for the calculation and any

[32] http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi
[33] https://eclipse.epochbio.com/login.asp
[34] http://www.premierbiosoft.com/netprimer/netprlaunch/netprlaunch.html

discordant curve was eliminated. The appropriate threshold to measure the cycle number (Ct) was selected and the iCycler iQ Detection (Bio-Rad, Munich, Germany) or the DNA Opticon® 2 (Biozym Diagnostik GmbH, Hess. Oldendorf, Germany) softwares automatically calculated the slope by plotting the threshold cycle against the logarithmus of the starting concentration. The efficiency of the reaction was calculated using the formula $E = 10^{[-1/slope]}$ (for the iCycler) or $E = 10^{[slope]}$ (for the Opticon 2). The efficiency of each primer combination given in Table 35 (Appendix) in percentage form was calculated as follows E(%) = E x 50

3.3.4. qRT-PCR Protocol

Real-time PCR of all genes quantified using panels H and T was performed in an iCycler iQ™ Real-Time PCR Detection System (Bio-Rad, Munich, Germany); those using panel O were quantified in a DNA Opticon® 2 Continuous Fluorescence Detection System (Biozym Diagnostik GmbH, Hess. Oldendorf, Germany). Panel O contained the following cDNAs: basal ganglia, cerebellum, occipital cortex, retina, RPE, bladder, distal colon, heart, and lung. Panels H and T included the same cDNAs plus stomach. The detection of the amplified product was achieved by using SYBR Green I (Sigma-Aldrich Chemie GmbH, Munich, Germany) at a final concentration of 0.5x. PCR amplification was performed in a 25 µl reaction mixture containing 5 µl of cDNA template which was prepared by diluting the reverse-transcribed cDNA 1:5. This extra dilution was done in order to pipette bigger amounts and thus minimize variations in the template quantity pipetted. Templates were pipetted with the 5-100 µl Research® pro pipette (Eppendorf, Wesseling-Berzdorf, Germany) and the master mixes with the Multipette® (Eppendorf, Wesseling-Berzdorf, Germany).

The 25 µl reaction mix included 1x PCR buffer (10x PCR buffer: 500 mM KCl, 100 mM Tris-HCl - pH 8.3), 200 nM of each primer, 0.5 µl of *Taq* polymerase, 100 µM of each dATP, dCTP, dGTP, 200 µM dUTP, 10 nM FITC (Bio-Rad, Munich, Germany), 0.5x SYBR Green I, and between 2 and 3 mM $MgCl_2$. All components were pipetted on ice and the reaction was transferred from the ice directly to the pre-heated block at 94°C. After denaturation at 94°C for 2 min the samples were amplified by repeating 40 times a cycle consisting of a denaturation step at 94°C for 30 sec, annealing at the optimized temperature for 30 sec (Table 35, Appendix), and an 8 sec extension at the optimal temperature for each primer (Table 35, Appendix). The melting curve analysis done after PCR amplification consisted of a 1 min incubation at 94°C, followed by a reduction of the temperature to the annealing temperature which was held for 1 min, and step-wise heating of the probes in intervals of 0.5°C.

3.3.5. Data analysis

The iCycler iQ Detection (Bio-Rad, Munich, Germany) and DNA Opticon® 2 (Biozym Diagnostik GmbH, Hess. Oldendorf, Germany) softwares together with an in-house spreadsheet based on the model published by Vandesompele et al. (2002) were used for the analysis of the results. First, the amplification curve for each sample was baseline-adjusted but not curve-fitted. Then, both the

amplification and melting curve of each sample were analyzed and outliers were eliminated from the data set. For each amplification plot, a threshold cycle (Ct) value was automatically calculated by the software and the Ct values were exported to the spreadsheet. After entering the efficiency of each PCR reaction the macros that were included in the spreadsheet facilitated the calculations. In short, the Ct value of each sample was transformed to a quantity using the formula $E^{(\text{min Ct of panel - Ct of interest})}$, where 'Ct of interest' indicates the cycle at which the template of interest trespasses the threshold and 'min Ct' refers to the lowest Ct value obtained for any of the samples included in the panel. The values for each gene of interest (GOI) were then normalized by dividing the GOI value through the normalization factor (NF) for the corresponding tissue. The NF factor was obtained by calculating the geometric mean of the values obtained for six housekeeping genes: ß-actin (ACTB), beta-2-microglobulin (B2M), ribosomal protein L13a (RPL13A), succinate dehydrogenase complex, subunit A (SDHA), hypoxantine phosphoribosyl-transferase I (HPRTI), beta-glucuronidase (GUS), and TATA box binding protein (TBP). The transformed values for each tissue were then averaged and the standard deviation for each gene of interest (SD GOI) was calculated using the formula:

$$SD\ GOI_{norm} = GOI_{norm} \cdot \sqrt{\left(\frac{SD\ NF_n}{NF_n}\right)^2 + \left(\frac{SD\ GOI}{GOI}\right)^2}$$

The standard deviation of the normalization factor based on *n* housekeepers (HKG_n) was calculated using the formula:

$$SD\ NF_n = NF_n \cdot \sqrt{\left(\frac{SD\ HKG_1}{n \cdot HKG_1}\right)^2 + \left(\frac{SD\ HKG_2}{n \cdot HKG_2}\right)^2 + \left(\frac{SD\ HKG_n}{n \cdot HKG_n}\right)^2}$$

In the last step, the expression values for the tissues were re-calculated so that the tissue with the least expression received a value of one and the expression of the other tissues were displayed as fold changes relative to this tissue. The relative expression was then plotted using the Excel chart function (Microsoft, Unterschleissheim, Germany).

3.4. Agarose gel electrophoresis

The electrophoretic separation of nucleic acids was done using a 1% agarose/ethidium bromide gel. For special applications concentrations of 0.7% up to 2.5% were used. The used buffer was 1x TBE (89 mM Tris-HCl, 89 mM borate acid, 2 mM Na_2EDTA - pH 8.3). Prior to loading, the DNA was mixed with a 6x gel-loading buffer (0.25% bromophenol blue, 40% sucrose) to an approximate final concentration of 1x. The electrophoresis was done at 140 V and DNA fragments were visualized with ultraviolet light. The size of the DNA fragments was estimated by comparison with the 1 kb Plus DNA Ladder™ (Invitrogen, Karlsruhe, Germany).

3.5. Purification of PCR products

To isolate DNA fragments from agarose gels the NucleoSpin® Extract kit (Macherey-Nagel, Düren, Germany) was used following the instructions of the manufacturer.

3.6. Cloning of PCR products

The fact that most polymerases add an adenine overhang at the end of the amplicon was used to clone the products in a vector with thymine overhangs. Two different commercial kits based on this principle were used. Cloning with the TA Cloning® Kit (Invitrogen, Karlsruhe, Germany) was done in a final volume of 10 µl with 1 µl T4 ligase, 1 µl buffer, 2 µl pCRTM2.1 vector, and desired amount of PCR product; the ligation occurred overnight at 14°C. The reaction with the pGEM®-T Vector Systems (Promega, Mannheim, Germany) included 5 µl of the 2X rapid ligation buffer, 1 µl T4 DNA ligase, 1 µl pGEM®-T vector, x µl of the PCR product, and dd H_2O to a final volume of 10 µl. The ligation was accomplished after 1 hr incubation at room temperature. The recombinant plasmid was then transformed into *E. coli* XL1-blue electrocompetent cells.

3.7. Sequencing of DNA sequences

All sequencing reactions were prepared with ABI PRISM® BigDye™ Terminators kit (Applied Biosystems, Weiterstadt, Germany) and run on the automated single-capillary genetic analyzer ABI PRISM® 310 Genetic Analyzer (Applied Biosystems, Weiterstadt, Germany). To prepare PCR products for sequencing an aliquot of approximately 1.5 µl (depending on the amount of product) was treated with shrimp alkaline phosphatase (SB, Cleveland, USA) and exonuclease I (USB, Cleveland, USA) in a final volume of 10 µl. The 10 µl sequencing reaction was set by mixing 5 µl of the treated product, 2 µl of BigDye mix, and 1 µl of 10 µM forward or reverse primer. In case the product to be sequenced had been purified from a gel, no digestion was necessary and 7 µl of the purified product were mixed with the same amount of BigDye and primer. In case plasmid DNA was to be sequenced the template quantity was reduced to 2 µl and the BigDye increased to 3 µl per reaction. After amplification of the product following the cycling conditions of the manufacturer, the product was purified by ethanol precipitation with 0.1 vol 3 M sodium acetate (pH 4.6) and 2.5 vol 100% ethanol. The sequences were viewed and analyzed using the Chromas software (Technelysium Pty Ltd, Helensvale, Australia).

3.8. Rapid amplification of cDNA ends (RACE)

3.8.1. 5'-RACE

Amplification of the 5'-end of a gene was achieved by using the 5'-RACE system (Invitrogen, Karlsruhe, Germany). In short, 1 µg of RNA was reverse transcribed using 2.5 pmoles of a gene-specific primer following the first-strand cDNA synthesis protocol (Materials and Methods 2.5). This product was then purified using the NucleoSpin® Extract Kit (Macherey-Nagel) following the manufacturer's instructions. An oligo dC anchor sequence was then added to the 5'-end of the cDNA by mixing 10 µl of the purified product with 5.0 µl 5x tailing buffer and 2.5 µl 2 mM dCT in a final

volume of 24 µl. This was incubated for 3 min at 94°C, chilled, and supplemented with 1 µl terminal deoxynucleotide transferase. After incubation at 37°C for 10 min and again for 10 min at 65°C, 5 µl of the tailed cDNA were amplified in a 50 µl PCR using a second gene-specific primer and a deoxyinosine-containing anchor primer (AAP, Table 35, Appendix). To reduce unspecific products a 1:100 dilution of the obtained product was then amplified with a second gene-specific primer and the AUAP primer (Table 35, Appendix). To facilitate the recognition of specific products negative control reactions were simultaneously run. They consisted of a reaction with cDNA which had not been tailed and a reaction including tailed cDNA but only with the AAP primer. All products were loaded on a gel, specific bands were excised, purified, cloned, and sequenced.

3.8.2. RNA ligase mediated rapid amplification of cDNA ends (RLM-RACE)

To identify the 5'-end of full-length transcripts, the FirstChoice™ RLM-RACE system was used (AMS Biotechnology GmbH, Wiesbaden, Germany). Following the manufacturer's instructions 10 µg of total retina RNA were treated with 2 µl of calf intestinal phosphatase (CIP) in a final volume of 20 µl containing 1x CIP buffer in order to remove the free 5'-phosphates present in all transcripts lacking the cap structure. The RNA was then purified by phenol:chloroform extraction and resuspended in 11 µl of H_2O. Half of the aliquot was added to a mix containing 5 µl tobacco acid pyrophosphatase (TAP), 1 µl of 10x TAP buffer and 2 µl of H_2O. The removal of the cap structure, which is found only in full-length mRNA, proceeded for 1 hr at 37°C. Afterwards, the 5' RACE adapter was ligated to the decapped transcript in a final volume of 10 µl containing 1x RNA ligase buffer, 1 µl of 5' RACE adapter, 5 U of T4 RNA ligase, and 2 µl of CIP/TAP-treated RNA. A minus-TAP control was also simultaneously started in order to identify and filter out unspecific products. Both the TAP+ and TAP- RNAs were reverse transcribed at 47°C for 1 hr in a 20 µl reaction containing 2 µl RNA, 4 µl dNTP mix, 2 µl random decamers, 2 µl 10x RT buffer, 1 µl RNase inhibitor, and 1 µl MMLV reverse transcriptase. This stock of cDNA was kept at -20°C and used to clone the 5'-ends of different genes.

The 5'-end of a specific gene was amplified by PCR using the usual PCR reagents, the previously prepared cDNAs (TAP+ and TAP-), a gene specific primer and the 5' RACE outer primer (Table 35, Appendix). If necessary, a nested PCR was also done using a second gene-specific primer and the 5' RACE inner primer.

3.8.3. Marathon-Ready™ cDNA

Marathon-Ready™ cDNAs (BD Biosciences Clontech, Heidelberg, Germany) are premade 'libraries' of adaptor-ligated double stranded cDNA. They can therefore be used as templates in amplifications to isolate the 5'-end of a gene using only an adaptor and a gene-specific primer. The human retina Marathon-Ready cDNA was used to clone the 5'-end of C1orf32 by amplification with primers A166F4 and AP1 (Table 35, Appendix) using the Advantage polymerase (BD Biosciences Clontech, Heidelberg, Germany) at an annealing temperature of 58°. Since many products were generated, a 1:100 dilution of the primary PCR product was amplified in a nested PCR with primers A166F6/AP2

(Table 36 and 35, Appendix) followed by a third PCR with primers A166F5 and AP2 (Table 35, Appendix) using a range of different $MgCl_2$ and formamide concentrations. The controls included water samples as well as reactions were only the AP2 primer was added to the cDNA.

3.9. Restriction enzyme analysis

Restriction enzyme analyses were done in order to investigate the frequency of polymorphisms and to prepare samples for Southern blot analysis. The amount of enzyme and conditions used are in accordance with the manufacturer's instructions (New England Biolabs, Frankfurt am Main, Germany). The reaction mix included the 1x reaction buffer, 1-4 U of enzyme for each µg of DNA, plus some additives for certain enzymes. The incubation temperature depended on the enzyme.

3.10. Single-strand conformational polymorphism analysis (SSCP)

Sequence variation frequencies were assessed by single-strand conformational polymorphism analysis (SSCP) (Orita et al. 1989). The PCR reaction was set up as explained in section 3.2, but with the addition of 0.1 µl a^{32}P-dCTP. If the amplicon was longer than 300 bp, a 5 µl aliquot of the product was digested in a final volume of 20 µl with the appropriate enzyme and reagents. Previous to loading, 5 µl of the digestion reaction or 3.5 µl of the PCR product were mixed with 1 vol of PAA-loading buffer (95% formamide, 5 mM NaOH, 0.1% bromophenolblue, and 0.1% xylencyanol) and denatured at 100°C for 5 min. Only 3.5 µl of sample were then loaded on a non-denaturing, glycerol-containing 6% PAA-gel made by mixing 48.5 ml H_2O, 10.5 ml Rotiphorese® Gel 40 (37.5:1) (Carl Roth GmbH & Co. KG, Karslruhe, Germany), 4.0 ml 85% glycerol, 7.0 ml 5x TBE and addition of 90 µl TEMED (Sigma-Aldrich Chemie GmbH, Munich, Germany) and 180 µl 40% APS (Sigma-Aldrich Chemie Gmbh, Munich, Germany). The gels were run in 0.5x TBE at 25 W for 260-500 min at 4°C. The gels were then transferred to Whatmann 3 mm filter paper and dried. Autoradiography was performed using X-ray Retina film (Fotochemische Werke GmbH, Berlin, Germany).

3.11. Southern blot analysis

Genomic DNA (10 µg) was digested with the selected enzymes (*EcoRI*, *HincII*, *PstI*) and appropriate buffers in a final volume of 60 µl. After the digestion was complete (tested by analysis of 0.1 vol of the reaction in a 0.6% gel) the remaineder of the reactions were loaded in a 0.6% gel without ethidium bromide and electrophoretically separated at 35 V during 20 hrs. After staining of the gel with ethidium bromide (Applichem, Darmstadt, Germany) and photography with a ruler, the gel was washed twice for 10 min in 0.25 M HCl in order to digest large fragments. The DNA fragments were then denatured twice for 10 min in an alkaline solution (0.5 M NaOH, 1.5 M NaCl) and subsequently neutralized in 25 mM Na_3PO_4, pH 6.5.

The digested and denatured DNA was then transferred from the gel to the Nylon Hybond N^+ membrane (Amersham Biosciences, Freiburg, Germany) by capillary blotting with 10x SSC (1.5 M

NaCl, 150 mM NaCitrate - pH 7.0). After marking the position of the wells on the membrane with a pencil, the membrane was UV-crosslinked so that the ssDNA stayed irreversibly bound. The hybridization of the membrane was done with probes prepared in the same way as explained in 2.4.2.

3.12. Virtual Northern blot analysis

Full-length cDNAs were generated from 2 to 4 µg of total RNA from bladder, heart, lung, skeletal muscle, retina, brain, and RPE. They were subjected to SMART cDNA synthesis following manufacturer's instructions (BD Biosciences Clontech, Heidelberg, Germany). This procedure relies on the fact that when the SuperScript II™ (Invitrogen, Karlsruhe, Germany) reaches the 5'-end of the template, its terminal transferase activity adds additional cytosines to the 3'-end of the first-strand cDNA. The additional primer included in the reaction hybridizes with the terminal tail, and the transcriptase switches strands and incorporates the primer sequence into the cDNA. Briefly, RNA was mixed with 1 µl of the CDSIII/3' PCR primer (10 µM), 1 µl of the SMARTIII oligo (10 µM), and 1 µl of dNTPs (10 mM each) and incubated at 72°C for 2 min and then immediately chilled. The reaction continued with the addition of 2 µl 5x First-Strand Buffer (250 mM Tris-HCl - pH 8.3, 375 mM KCl, 15 mM $MgCl_2$), 1 µl 0.1 M DTT, and 200 U of SuperScript II™ (Invitrogen, Karlsruhe, Germany). The first-strand synthesis was then allowed to proceed for 1 hr at 42°C.

The first-strand cDNAs were amplified in a 25 µl (for optimization) or in 100 µl (amplification for blots) reaction mixture. The PCR cycling profile included a denaturation at 94°C for 2 min followed by the basic protocol of denaturation at 94°C for 15 sec, annealing at 64° for 30 sec, and elongation at 68°C for 6 min 30 sec. The optimal quantity of template and number of cycles was determined individually for each tissue. During optimization of the amplifications both the Advantage *Taq* (BD Biosciences Clontech, Heidelberg, Germany) and the KOD HiFi *Taq* (Merck Biosciences GmbH, Schwalbach, Germany) were used. For its better performance the latter was chosen for the preparative PCRs. The reaction mix included 1x PCR buffer (10x PCR buffer: 1.2 M Tris-HCl, 100 mM KCl, 60 mM (NH4)2SO4, 1% Triton X-100, 0.01% BSA - pH 8.0), 1x dNTP mix, 200 nM 5' PCR primer (Table 35, Appendix), 200 nM CDSIII/3' PCR primer, 1 mM $MgCl_2$, and 10 U KOD polymerase in a final volume of 100 µl.

Five microliters of the amplified cDNA products were resolved on a 0.7% TBE/agarose/Ethidium bromide gel and subjected to Southern blotting and probe hybridization as described above (Material and Methods 2.4.2 and 3.11). To increase the sensitivity and shorten the exposure times, most of the filters were exposed to storage phosphor screens (Amersham Biosciences, Freiburg, Germany) and the signals were visualized using the Typhoon™ 9200 Imager (Amersham Biosciences, Freiburg, Germany). The rest of the hybridized membranes were exposed to X-ray Retina film (Fotochemische Werke GmbH, Berlin, Germany) and incubated at -80°C.

3.13. Amplification of chromosome panels

To determine the chromosomal location of a specific DNA fragment, a PCR was carried out using as template the different hybrid DNAs included in the Coriell human/rodent chromosome mapping panel #2, version 3 (NIGMS Human Genetic Mutant Cell Repository, Coriell Institute for Medical Research, Camden, USA). This panel consists of 24 rodent / human hybrid cell lines each of which has retained a single human chromosome. The majority of the chromosomes are found in the Chinese hamster cell line, only chromosomes 1, 16, 17, 20, and 21 are found as hybrids in the mouse cell line. The panel also includes the three parental DNAs from human line IMR91, mouse line 3T6, and Chinese hamster line RJK88.

4. cDNA libraries

4.1. Screening of cDNA pooled libraries

For selected genes, a series of cDNA pools provided by the RZPD[35] center (Berlin, Germany) were screened by PCR (Table 3).

Table 3 List of the screened libraries and their corresponding library number

Lib. no.	Description	Lib. no.	Description
409	Human brain corpus callosum	578	Human brain amygdala
414	Human caudate nucleus	578	Human spleen
415	Human fetal brain	584	Human pituitary gland
420	Human thymus	588	Human adult brain
424	Human pericardium	589	Human fetal kidney
440	Human osteoarthritic	595	Human brain thalamus
514	Human pancreas	597	Human spleen
518	Human control	636	Human scalp
519	Human control	727	Human breast cancer cell line
569	Human hippocampus	761	Human brain amygdala tissue

If a product could be amplified from one of the pools, the filters containing arrayed clones from the library were screened. The screening was done using a radiolabeled probe using the same protocol as for Northern blot (Material and Methods 2.4.2 and 2.4.3). Since each clone is arrayed twice on each filter, two signals per positive clone are observed. The position of the signals in the grid was used to identify the exact desired clone. Clones were ordered from RZPD (Berlin, Germany) by submission of the X and Y coordinates of the signal.

4.2. Identification of cDNA clones of interest by library screening

4.2.1. Library plating

Clones containing the sequence of genes investigated in this project were isolated from a number of phage libraries (Table 4) by screening of membranes containing replicas of the clones. The replicas were obtained by making lifts of LB plates where the insert-containing phages grow. Phage libraries

[35] http://www.rzpd.de

are grown by infection of a host cell with the insert-containing phages. For all libraries constructed in the ?TriplEx2 vector the XL1-Blue strain was used; Y1090r $^-$ host cells were used for the Nathan's library (Table 4).

Table 4 Available cDNA libraries

Tissue	Vector	Constructed by	Laboratory ID
Adult retina	?gt10 HRET	J. Nathans (Johns Hopkins University, Baltimore, USA)	Nathans
Adult retina	?TriplEx2	Andrea Gehrig (University of Wuerzburg, Germany)	Andrea
Adult retina	?TriplEx2	Jelena Stojic (University of Wuerzburg, Germany)	Jelena
Adult retina	?TriplEx2	Claudia Berger (University of Wuerzburg, Germany)	C1F1, C1F2, C2F3*
Adult retina	?TriplEx2	DKFZ (Heidelberg, Germany) and Claudia Berger (University of Wuerzburg, Germany)	DKFZ1, DKFZ2, DKFZ3, DKFZ4*
Fetal eye	?gt11	Library donated by Dr. David Kurmit, Howard Hughes Medical Institute, Ann Arbor, USA. The library was constructed from an 18-week fetus.	Fetal eye

*Before packing the fragments were size-selected. Each fraction was packed independently and therefore there are several lab IDs, each identifying one of the fractions

Before a library was plated the titer of the library was determined by mixing 10 µl serial dilutions of library aliquots with 600 µl of host cells freshly grown in $MgSO_4$-containing LB (20 g/l peptone, 10 g/l yeast extract, 20 g/l NaCl, 20 mM $MgSO_4$, and 0.2% maltose). After incubating 15 min at 37°C, 10 ml of molten Top-Agar (20 g/l peptone, 10 g/l yeast extract, 20 g/l NaCl, and 7.2 g/l agar) were added and the mix was plated on an LB plate (20 g/l peptone, 10 g/l yeast extract, 20 g/l NaCl, and 15 g/l agar). After overnight incubation at 37°C the number of plate forming units (pfu) in each plate was counted and the titer (expressed as pfu/ml) was calculated. For each library approximately $1x10^6$ clones were plated on six Bio-Assay dishes (NUNC, Wiesbaden, Germany) as described above but with the exception that 100 µl of diluted library aliquot was mixed with 2.1 ml of host cells. The amount of time needed until the plaques reached the optimal size was carefully monitored in order to interrupt the incubation at the point when the plaques had reached about 2 mm of diameter but were still not confluent (usually between 6 and 12 hrs).

4.2.2. Generation of phage replicas

A double set of membranes from each plate were obtained by duplicate lifting of each plate using the Nylon Hybond N^+ membrane (Amersham Biosciences, Freiburg, Germany). After the lift was made, the membrane was placed side face-down on top of a filter paper soaked with denaturating solution (0.5 M NaOH, 1.5 M NaCl) for 2 min and was subsequently neutralized in 25 mM Na_3PO_4, pH 6.5.

4.2.3. Identification and isolation of a specific clone from a phage library

To isolate a specific clone the duplicate library membranes were hybridized with a radio-labeled probe (Material and Methods 2.4.2 and 2.4.3). In order to purify the desired clone, a small sector of the plate corresponding to the positive signal was cut out and stored in 1 ml of SM buffer (5.8 g NaCl, 2 g

$MgSO_4$, 25 ml 1 M Tris-HCl - pH 7.5, 2.5 ml 2% gelatine in a final volume of 1000 ml H_2O). After a 1 hr incubation, serial dilutions of the phage-containing supernatant were made and plated out in Petri dishes (Ø 135 mM) as explained above in order to determine the optimal concentration so that approximately 200 plaques would be present in each dish. Once the right density of pfu was obtained, a lift was made (Material and Methods 4.2.2) using the Colony/Plaque Screen™ (DuPont, Bad Homburg, Germany) and this membrane was hybridized with the same probe used in the first screening. Positive plaques were visualized by autoradiography, picked, and stored in SM buffer.

4.2.4. Conversion of phages

The ?TriplEx2 vector can be converted via Cre-lox recombination into the pTriplEx2 plasmid. To achieve this, a single colony from an *E. coli* BM25.8 stock plate was inoculated in 10 ml of LB broth and incubated at 31°C overnight with shaking. Previous to the addition of the phage solution, 100 µl of 1 M $MgCl_2$ were added to the culture. For the conversion, 200 µl of the cell culture was combined with 150 µl of the eluted positive plaque and the mix was incubated at 31°C for 30 min without shaking. After addition of 400 µl of LB broth it was again incubated at 31°C for an additional hour with shaking. In order to isolate a plasmid containing colony, 1-10 µl of infected cell suspension were spread on an LB/ampicillin plate.

4.3. Construction and sequencing of a suppression subtracted hybridization cDNA library (SSH)

Using the Oligotex mRNA Purification System (Qiagen, Hilden, Germany) PolyA$^+$ RNA was isolated from total adult human retina, liver and kidney RNA. The adult human retina SSH (retSSH) cDNA library was constructed using the PCR-Select™ cDNA Subtraction Kit (BD Biosciences Clontech, Heidelberg, Germany) following the manufacturer's protocol. This work was done by Andrea Gehrig, Institute of Human Genetics, University of Wuerzburg, Germany. In short, 50 ng of retina PolyA$^+$, and 100 ng of kidney and liver PolyA$^+$ were reverse transcribed using the SMARTII Oligonucleotide and cDNA synthesis primer (CDS) (BD Biosciences Clontech, Heidelberg, Germany). In order to generate double strand cDNA, first-strand cDNAs were amplified by long-distance (LD) PCR which was run for 15 cycles for liver and 18 cycles for retina and kidney. This was followed by *RsaI* digestion of all samples and ligation of equal aliquots of retina cDNA to either adaptor 1 or 2R. In order to normalize and enrich for sequences differentially expressed in retina two rounds of hybridization were done. In the first, an excess of denatured driver cDNA (mix of liver and kidney cDNA) was added to each denatured aliquot of anchor-containing retina. In the second hybridization the two primary hybridization samples were mixed together and a small amount of denatured driver cDNA was added. Thus, adaptor-ligated retina cDNA (tester) was mixed in a ratio of 1:35 with driver cDNA. After the hybridizations, the normalized population of retina-specific cDNAs with different adaptors on each end was amplified in two rounds of PCR. In the first PCR the sample was subjected to 27 cycles of amplification with PCR primer 1, the product was diluted tenfold and amplified 12 cycles with nested PCR primer 1 and 2R. To control the efficiency of the subtraction and normalization, a parallel reaction

with a retina aliquot which was not hybridized with driver cDNA was also done. The quality of the generated PCR products was verified by analyzing the products of the subtracted and un-subtracted retina from both the first and second PCR on a 1x TAE 2% agarose gel containing ethidium bromide (Applichem, Darmstadt, Germany). This gel was blotted and hybridized with radio-labeled probes to control the amount of specific transcripts present in the samples. The reduction of ubiquitous genes was tested by hybridization of a glyceraldehyde-3-phosphate dehydrogenase probe. An enrichment of retina-specific genes was corroborated by hybridization with probes amplified from the retinoschisis and rhodopsin genes. The last step in the library construction process consisted in the ligation of 2 µl of subtracted retina PCR product in the pCR®2.1-TOPO TA cloning vector (Invitrogen, Karlsruhe, Germany). After an overnight ligation at 14°C, 2 µl were transformed in DH5a™-T1® chemically competent cells.

An aliquot of transformed cells containing the retSSH PCR products was plated out on LB-Amp plates and grown overnight at 37°C. Colonies were randomly picked, resuspended in 100 µl of LB-Amp, and after overnight incubation at 37 °C PCR-amplified with primers M13F and M13R (Table 35, Appendix) to check if they contained an insert. The PCR product of all positive clones was then sequenced using either primer PCR-r or PCR-f (Table 35, Appendix). Sequencing was carried out with primer walking technology using either an ABI 310 automated sequencer and the ABI PRISM Ready Reaction Sequencing Kit (Perkin Elmer, Norwalk, USA) or a Beckman CEQ 2000 sequencer with the corresponding Dye Terminator Cycle Sequencing with Quick Start Kit (Beckman Coulter, Fullerton, USA).

V Results

1. Analysis and data-mining of the UniGene dataset

1.1. Pilot study of the UniGene cluster composition of genes associated with hereditary retinal diseases

The use of expressed sequence tags (ESTs), which are short single-pass sequences of 200 to 500 nucleotides, to investigate transcripts expressed in a certain tissue was introduced in 1992 by Okubo et al. The creation of the UniGene database[36] enhanced the use of ESTs by grouping all the sequences which derived from a gene in so-called clusters. In theory, the origin of the ESTs representing a gene provide a coarse first impression of a gene's expression pattern. We decided to investigate if the cluster composition can be used to predict potential retinal genes. To achieve this, we surveyed the information available for genes known to be involved in retinal disease whose expression had been reported. This analysis also served to evaluate the involvement of retina-specific or preferentially expressed genes in disease.

As of January 2000, the chromosomal localizations of a total of 118 different genes involved in retinal diseases were publicly available in the RetNet database[37]. From 56 identified and fully characterized genes, 18 cause syndromic diseases (e.g. Usher's syndrome, Refsum disease, papillorenal syndrome) and were not included in the evaluation. The red and green cone pigment genes were also omitted, leaving a total of 36 genes for the analysis. The composition of the UniGene clusters of these genes was investigated by comparing the number of eye ESTs to the number of ESTs from other tissue sources (Table 5). Seven of the disease-associated genes were represented by eye ESTs only, for 18 genes the ESTs derived from cDNA libraries of eye and other tissues, and 11 clusters did not contain any eye EST. This information was compared with the reported expression for each gene which was established using RT-PCR, Northern blot, or *in-situ* hybridization. Almost half of the genes (15 from 36) were reported to be expressed exclusively in retina. From the remaining genes, 10 are expressed in retina and other tissues, four were not detected in retinal tissue with the methods used, and seven are ubiquitously expressed (Table 5).

A correlation between the EST and the actual expression profile was evident and led us to define five categories (Table 6). The classification was based on the proportion of eye ESTs in an EST cluster. Category I and II encompass genes with an absolute ratio of 1. The genes in category I are represented by three or more eye ESTs, whereas category II consists of genes with only one or two eye ESTs. The third and fourth categories include genes with ESTs derived from eye and other tissues. Whereas in category III more than a third of the ESTs originate from eye cDNA libraries, in category IV the proportion of eye ESTs ranged between a third and a tenth of the total ESTs. Category V includes the genes represented by eye ESTs in a proportion smaller than a tenth of the total ESTs.

[36] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene
[37] http://www.sph.uth.tmc.edu/Retnet/

The rhodopsin kinase gene (GenBank acc. no. U63973), which has been shown to cause congenital stationary night blindness Oguchi type, was also included in category V even though no EST sequences representing this gene were reported.

Table 5 Cluster composition and expression profiling of known retinal disease genes

	Gene	Disease	Nr. of ESTs (eye + other)	Absolute ratio	Expression profile (*)	Reference for expression profile
Category I	RHO	Autosomal dominant or recessive RP (RP4); congenital stationary night blindness (CSNB)	81+0	1	Retina (N)	Nathans and Hogness, 1984
	GNAT1	Dominant CSNB	44+0	1	Rod photoreceptor cells (N)	Fong, 1992
	GUCY2D	Leber congenital amaurosis (LCA), dominant cone-rod dystrophy (CRD) type 6	3+0	1	Retinal outer nuclear layer (ISH)	Shyjan et al. 1992
Category II	RP1	Autosomal dominant RP	2+0	1	Retina (N)	Pierce et al. 1999
	CRX	Cone-rod dystrophy-2, RP, LCA	2+0	1	Retina (N)	Furukawa et al. 1997
	CACNA1F	X-linked CSNB	2+0	1	Retina, skeletal muscle, kidney, pancreas (N)	Bech-Hansen et al. 1998
	CNGA3	Recessive achromatopsia	1+0	1	Testis, kidney and heart (N, RT)	Kohl et al. 1998
Category III	PDE6G	Mouse recessive retinal degeneration	78+17	0.82	Retina (N)	Tuteja et al. 1990
	SAG	Recessive Oguchi disease, recessive RP	33+3	0.92	Retina, pineal gland (N, W)	Craft et al. 1990
	PDE6A	Recessive RP	19+2	0.90	Retina (N)	Pittler et al. 1992
	ABCA4	Juvenile and late onset recessive Stargardt disease; recessive macular degeneration (MD); recessive RP; recessive fundus flavimaculatus; recessive combined RP and CRD	16+21	0.43	Retina (ISH)	Allikmets et al. 1997
	ROM1	Dominant RP; digenic RP	15+15	0.50	Retina (N)	Bascom et al. 1992
	NRL	Autosomal dominant RP	14+2	0.88	Neuronal cells of retina (N)	Swaroop et al. 1992
	RLBP1	Recessive RP; recessive Bothnia dystrophy; recessive retinitis punctata albescens	6+3	0.67	Retina and pineal gland (N)	Crabb et al. 1988
	RDS	Dominant RP; dominant MD; digenic RP with ROM1; dominant adult vitelliform MD	5+7	0.42	Retina (n.k.)	Travis et al. 1991
	RS1	X-linked, juvenile retinoschisis	5+2	0.71	Retina (N)	Sauer et al. 1997
Cat. IV	NDP	Norrie disease; familial exudative vitreoretino-pathy; Coats disease	2+12	0.14	Fetal and adult brain, fetal lung, fetal eye (RT)	Chen et al. 1993
	PDE6B	Recessive RP; dominant CSNB	2+10	0.17	Retina, smaller transcript in brain (N)	Collins et al. 1992
	VMD2	Atypical vitelliform dominant MD	2+10	0.17	RPE. Lower expression in retina, brain and testis (N, RT)	Marquardt et al. 1998 ; Petrukhin et al. 1998
Category V	TULP1	Recessive RP	1+4	0.20	Retina (N)	North et al. 1997
	EFEMP1	Dominant radial, macular drusen; dominant Doyne honeycomb retinal dege- neration (Malattia Leventinese)	13+198	0.06	Ubiquitious (N, RT)	Stone et al. 1999
	TIMP3	Dominant Sorsby's fundus dystrophy	11+1143	0.01	Ubiquitious (N)	Silbiger et al. 1994
	PGK1	RP with myopathy	9+385	0.02	Ubiquitious (N)	Michelson et al. 1983
	RDH5	Recessive fundus albipunctatus	1+28	0.03	Ubiquitious (N)	Wang et al. 1999
	CNGA1	Recessive RP	1+17	0.06	Retina, heart, kidney (N)	Ahmad et al. 1990
	RPE65	Recessive LCA; recessive RP	0+1	-	RPE (n.k.)	Bavik et al. 1993
	RB1	Dominant germline or somatic retinoblastoma; benign retinoma; pinealoma; osteogenic sarcoma	0+75	-	Fetal retina and placenta, tumour tissues (N)	Lee et al. 1987
	PAX2	Dominant renal-coloboma syndrome	0+19	-	Developing eye, ear and kidneys (ISH)	Terzic et al. 1998
	OAT	Recessive gyrate atrophy	0+153	-	Ubiquitious (N)	Mitchell et al. 1988
	RPGR	X-linked RP, X-linked CSNB	0+15	-	Ubiquitious (N)	Meindl et al. 1996
	RP2	X-linked RP	0+15	-	Ubiquitious (ISH)	Schwahn et al. 1998
	RBP4	Recessive RPE degeneration	0+138	-	Liver (n.k.)	D'Onofrio et al. 1985
	CHM	Choroideremia	0+11	-	Retina and brain with lower expression in liver, ciliary body (N)	Bernstein and Wong, 1998
	BCP	Dominant tritanopia	0+1	-	Cone photoreceptor cells (N)	Nathans et al. 1986
	TTPA	Recessive RP and/or recessive or dominant ataxia	0+1	-	Liver (N). Lower expression in brain, spleen, lung, kidney (ISH)	Arita et al. 1995 Hosomi et al. 1998
	RHOK	Recessive CSNB, Oguchi type	0+0	-	Retina (N)	Lorenz et al. 1991

(*) Abbreviations: CNS, central nervous system; ISH, *in-situ* hybridization; N, northern blot; n.k., not known; RT, RT-PCR; W, western blot

Table 6 Summary of cluster composition and expression of retinal disease genes

Category	Description	Average absolute ratio[a]	Number of disease genes	Genes expressed in retina only
I	Only retina ESTs, 3 or more	1.00	3	3
II	Only retina ESTs, 1-2	1.00	4	2
III	Retina ESTs / total ESTs: > 0.3	0.70	9	9[b]
IV	Retina ESTs / total ESTs: > 0.1 and < 0.3	0.17	4	1
V	No retina ESTs or retina ESTs / total ESTs < 0.1	0.01	16	2
	TOTAL		**36**	**17**

[a] Absolute ratio = nr. of eye ESTs / nr. of total ESTs
[b] Two of the genes are also expressed in pineal gland

The three genes which had an EST composition fulfilling the criteria for category I, RHO, GNAT1 and GUCY2D, are expressed only in retina. As expected, there is a high number of eye ESTs (81) for RHO, the predominant protein in the rod photoreceptors (Table 5). Category II includes four genes, of which RP1 and CRX display retina-specific expression (Table 5). The CACNA1F gene is found in retina and other tissues, whereas CNGA3 expression in retina could not be detected by Northern blot hybridization. Therefore, this gene does not seem to be expressed in retina at high levels. Nine genes were included in category III with an absolute ratio between eye ESTs and total ESTs ranging from 0.4 to 0.9, and an average score of 0.7. According to published data, seven genes are retina-specific; the S-antigen and RLBP1 genes are additionally expressed in the pinealocytes, cells which are evolutionarily related to photoreceptor cells (Vollrath et al. 1985). Four genes were grouped in category IV, with an average absolute ratio of 0.17. Here, only the RP14 gene shows retina-specific expression whereas the PDE6B gene, which is involved in the phototransduction cascade, is also expressed in brain as a smaller transcript. A total of 16 genes were classified in category V. The absolute ratio for the disease genes in this category ranges from 0.0 to 0.06. This heterogeneous category includes genes with a high number of reported ESTs (e.g. TIMP3, EFEMP1, and PGK1) and others with only one or no reported EST (e.g. BCP, TTPA, and RHOK). Only the rhodopsin kinase gene and the blue cone pigment genes show retina-specific expression, even though no eye EST is reported for these genes. Almost 50% of the genes (7) in category V are ubiquitously expressed.

The analysis of the cluster composition of genes involved in retina diseases and comparison with the reported expression revealed that there is some correlation between both. Nevertheless, there are also many exceptions. Based on the low number of retina-specific genes found in the fourth and fifth categories, it was decided to exclude clusters with a retina EST proportion lower than 30% in future analyses.

1.2. Data mining of the UniGene dataset

In order to systematically identify genes expressed preferentially in the retina an *in-silico* approach using the UniGene human dataset was selected since it could be demonstrated that there is some correlation between cluster composition and expression. A total of 6190 retina-specific or enriched clusters were retrieved from the UniGene build #113 (June 2000). The clusters were identified by querying using the term 'eye' and therefore they contained at least one EST from any of the following 16 cDNA libraries: lib.125 human eye (148 ESTs); lib.165 subtracted human retina (32 ESTs); lib.169 subtractive cDNA library ocular ciliary body (196 ESTs); lib. 177 Soares retina N2b4HR (1707 ESTs); lib.178 Soares retina N2b5HR (9169 ESTs); lib.190 human ocular ciliary body cDNA library 1 (1 EST); lib.191 human ocular ciliary body cDNA library 2 (1 EST); lib.221 human fovea cDNA (42 ESTs); lib.300 human retina (D. Swanson) (7 ESTs); lib.226 human retina cDNA randomly primed sublibrary (803 ESTs); lib.228 human retina cDNA Tsp509I-cleaved sublibrary (2079 ESTs); lib.277 Stratagene fetal retina (2422 ESTs); lib.313 retina II (1080 ESTs); lib.433 retina I (19 ESTs); lib.450 cornea I (17 ESTs) and lib.451 cornea II (13 ESTs). Since 97.9% of the ESTs (17,360/17,736) were sequenced from cDNA libraries that have been generated from RNA isolated from retinal tissue, the ESTs are referred to as retina ESTs hereafter.

To keep track of the large amount of collected information a database was designed using the Access platform (Microsoft, Unterschleissheim, Germany) where a record was created for each of the 6190 clusters. For the known genes only the gene name and Hs. number was recorded in the corresponding fields. For all other clusters the following fields were available: Hs. number, KIAA number, number of retina ESTs, total number of ESTs in the cluster, lab ID assigned to cluster, name of person who analyzed the cluster, chromosomal localization, number of the corresponding 5'- and 3'-'tentative human consensus' (THC) record, existence of repeat sequence in the THC, GenBank accession number of the genomic sequence (if available), and comments (Fig. 6).

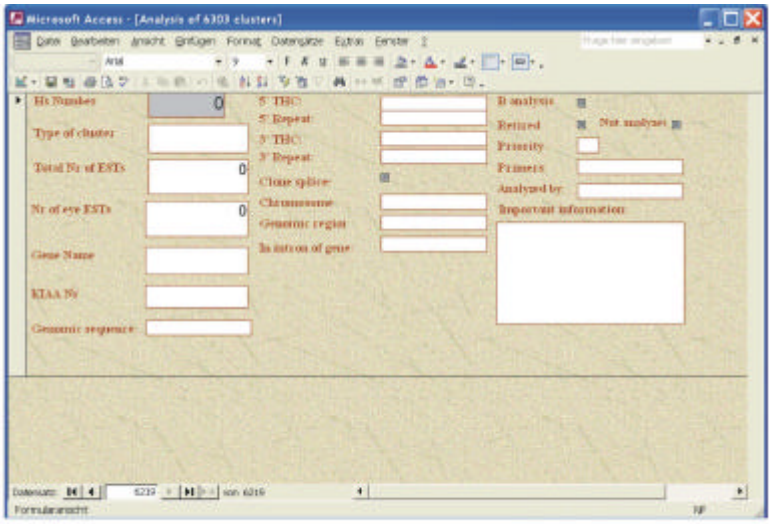


Fig. 6 Formulary of the UniGene database
To optimize administration of data generated in the UniGene project an Access database was created.
Information about the 6190 UniGene clusters containing retina ESTs was entered using this formulary.

1.3. Classification of retinal UniGene clusters

Analysis of the 6190 UniGene clusters revealed that more than one third (2201 clusters) represented previously identified and characterized human genes. They were therefore eliminated from further analyses. The remaining 3989 UniGene records had not been assigned to known genes but 834 entries included at least one full-length clone. The analysis of the 3898 yet unknown genes was set forth by assessment of the number of retina ESTs in each cluster. In order to investigate which cluster composition has the greatest probability of representing retina-specific genes, the 3989 entries were sorted according to their specificity for retina ESTs in eight categories. Two main categories which were defined as A: clusters composed only of retina ESTs and B: clusters which also contain sequence from other tissues were subdivided in four subcategories according to the number or retina ESTs.

Table 7 Classification of selected UniGene clusters containing retina ESTs

	Category A (retina EST-specific)		Subcategory	Category B (retina EST-enriched [a])		Subcategory
No. of retina ESTs in cluster	**No. of clusters**	**No. of clusters analyzed (%)**		**No. of clusters**	**No. of clusters analyzed (%)**	
1	435	22 (5.1)	A1	340	21 (6.2)	B1
2	122	21 (17.2)	A2	95	21 (22.1)	B2
3-5	70	22 (31.4)	A3	77	21 (27.3)	B3
>5	46	22 (47.8)	A4	56	30 (53.6)	B4
Total	**673**	**87 (12.9)**		**568**	**93 (16.4)**	

[a] A cluster is defined as retina EST-enriched if the no. of retina ESTs / no. of total ESTs in the cluster = 0.3

A total of 673 (16.9%) clusters were exclusively composed of retina ESTs (category A, Table 7). Of these, most (82.8%) were singleton retina ESTs (435, subcategory A1) or clusters with two ESTs (122, subcategory A2), whereas only 70 and 46 clusters were composed of 3–5 (subcategory A3) and more than 5 ESTs (subcategory A4), respectively (Table 7). Among the 3316 clusters with additional non-retina EST hits, 568 entries were considered retina EST-enriched with at least 30% of the ESTs derived from eye cDNA libraries (category B, Table 7). As in the retina-specific group A, the majority of clusters in category B (76.5%) contained either 1 (340, subcategory B1), or 2 retina ESTs (95, subcategory B2) (Table 7). Only 77 clusters included 3–5 retina ESTs (subcategory B3), and 56 sets had 6 or more retina ESTs (subcategory B4).

1.4. Phase I analysis of selected UniGene clusters

The identification of retina-preferential genes by using cluster composition as selection criteria was evaluated in a pilot that investigated the expression of 180 clusters representing all eight subcategories. The analysis included *in-silico* analysis of each cluster, design of primers, and expression analysis by RT-PCR.

The initial *in-silico* analysis included searches using representative 5'- and 3'-read EST sequences of each cluster in order to retrieve the corresponding THC sequences at TIGR[38]. In-house sequence alignments of all sequences were then carried out using the ClustalW 1.8[39]. Repeats in EST sequences were masked using the RepeatMasker[40] and sequence similarity searches against nucleotide sequence databases and unfinished high throughput genomic sequences (htgs) were carried out at NCBI[41]. On the basis of the consensus sequence, it was possible to design primers for 218 of the 341 analyzed EST clusters. Each of the clusters for which a primer was designed received a lab ID consisting of the letter A and a number, e.g. A001 (Table 8).

As of June 2000, alignment of the respective ESTs to genomic sequences identified 10 clusters representing mRNA species spliced at highly conserved exon-intron boundaries (Hs.28411, Hs.33791, Hs.60473, Hs.60563, Hs.64616, Hs.125827, Hs.146060, Hs.173105, Hs.175480, and Hs.275335). In six cases, two EST clusters were contained within the same genomic BAC/PAC clone, namely, A065 and A166, A067 and A174, A079 and A128, A105 and A177, A182 and A202, and A206 and A218.

To keep the cluster information updated, *in-silico* analyses for each of the 180 clusters was periodically repeated. In June, 2003 all the entries were revalidated in light of the most recent revisions of the public databases (e.g. genome browser at USCS[42]) and a summary of the current knowledge about each cluster was recorded (Table 8 and Table 9). Since UniGene has undergone several updates since the beginning of the study, the UniGene cluster numbers used in UniGene build #160 (February, 2003) have also been included.

[38] http://tigrblast.tigr.org/tgi/
[39] http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html
[40] http://repeatmasker. genome.washington.edu/cgi-bin/Repeat-Masker
[41] http://www.ncbi.nlm.nih.gov/blast/
[42] http://genome.ucsc.edu/cgi-bin/hgBlat?command=start

Table 8 Characteristics and expression of retina EST-specific clusters selected for expression analysis in UniGene Phase I

Lab ID	Original Hs. no.	Current Hs. no.[a] (EST acc. no.)	GenBank Acc. No.	Chromosomal localization [a]	No. of retina ESTs	Reported gene or full-length sequence [a]	In intron of: [a]	*In-vitro* expression[b]
A1								
A012	116914	116914	G65559	04p15.32	1		BCL2L13	O
A016	228657	228657	G65562	02q37.3	1	DKFZp667O1523	HDAC4	U
A033	98812	98812	G65567	10q24.32	1		SHFM3	U
A100	32524	32524	G65672	07p15.3	1			U
A103	32719	- (H37916)	G65673	03q11.2	1			R
A154	40388	40388	G65713	18q22.3	1			O
A155	40594	40594	G65714	12q21.2	1			R
A156	105309	105309	G65581	13q21.2	1		AK092024	U
A157	116913	116913	G65722	Xq25	1		ODZ1	O
A158	117158	151001	G65582	16q22.1	1	CIRH1A		O
A159	144156	144156	G65583	9q23	1			R
A160	174871	174871	G65584	12q21.2	1			U
A162	105308	105308	G65586	8p21.1	1		MGC45780	O
A164	275691	405040	G65587	n.a.	1			U
A165	237682	237682	G65588	5q21.1	1		CHD1	R
A166	237687	237687	G65589	01q24.1	1	C1orf32		R
A168	167937	167937	G65590	02q33.1	1		ORC2L	N
A176	175480	332256	G65596	04q34.2	1	WDR17		N
A177	174371	400579	G65597	01p31.2	1	DKFZp761D221		N
A178	174369	174369	G65598	08q22.3	1			O
A206	59830	59830	G65622	07p15.2	1			N
A216	40707	40707	G65630	02p22.2	1		AB037835	U
A2								
A011	60575	136037 & 442322	G65558	04p15.31	2		KCNIP4	N
A013	118681	118681	G65560	22q11.21	2			U
A015	269244	269244	G65561	17q25.3	2		KIAA1554	U
A017	268792	424926	G65563	01q42.3	2		ERO1LB	R
A018	269206	418395	G65564	05p14.2	2		CDH10	N
A085	40863	12513	G65662	08q23.1	2	LOC157627		N
A086	40918	40918	G65663	4q21.21	2			O
A087	269246	-	G65664	Xq26.3	2	Probably FHL1		U
A088	268813	418380	G65665	18q12.1	2			N
A089	269247	269247	G65666	10p15.3	2		AB023151	U
A112	261189	261189	G65716	12q23.3	2			R
A114	221447	221447	G65683	12q13.3	2		LOC56901	O
A152	32763	32763	G65721	04q32.1	2		GRIA2	N
A153	32856	32856	G65712	12q14.3	2			U
A161	269224	418391 & 231974	G65585	02p24.1	2			O
A169	275231	275231	G65591	16q23.1	2		ADAMTS18	R
A170	275211	337696	G65592	14q24.2	2	SLC8A3		N
A173	269280	-	G65593	16p13.3	2			O
A174	269238	269238	G65594	06q21	2		AF520419	R
A175	269221	418033	G65595	10p11.21	2		CFP1	O
A217	190309	-	G65631	17q25.3	2		EVER1	O
A3								
A002	32853	32853	G65550	05q22.3	5		KIAA1281	U
A005	269223	269223	G65552	04q31.3	5		BC014937	U
A006	191327	191327	G65553	04p13	3		KIAA1102	U
A043	32718	32718	G65571	0q32.3	4		FLJ20420	U
A091	59767	59767	G65668	2p12	4			O
A125	61094	61094	G65693	Xq23	3			R
A127	60545	60545	G65695	17q21.32	5		KIAA0924	O
A128	60673	60673	G65696	02q33.1	4	MPP4		R
A129	60473	60473	G65697	07p15.3	5	RFRP		R
A130	40513	40513	G65698	05q23.1	4		PTD002	R
A131	40567	40567	G65699	11q13.3	3	Probably TPC2		U
A134	60602	60602	G65701	07p21.2	5			O
A135	60740	425093	G65702	05q31.3	5			U
A140	60764	60764	G65704	03p12.1	3			N
A141	60781	60781	G65705	19q13.41	4			O
A143	108556	108556	G65717	05q14.1	5		MSH3	U
A144	110287	110287	G65707	01p31.1	3		LOC51086	O
A145	173105	194408	G65718	06q23.3	4	KIAA1244		U
A146	269265	117964	G65708	02p15	3	AK023367		U
A147	271791	271791	G65709	03q23	4			U
A149	171611	171611	G65710	15q21.1	3			O
A180	278418	222236	G65599	19q13.43	4	LOC116412		R
A4								
A001	32703	32703	G65549	09q22.3	6		KIAA1529	U
A004	40486	40486	G65551	01p36.23	8		CAMTA1	N
A007	59847	59847	G65554	12q13.12	17		FLJ34278	U
A008	33792	33792	G65555	12q21.31	19		FLJ21963	U
A009	271780	271780	G65556	04q26	18		CAMK2D	R
A045	40629	40629	G65573	09q22.23	16	FTP9Q22		O
A046	40814	40814	G65574	10q25.2	12			O
A047	40861	40861	G65575	10q21.1	8		PCDH15	N
A049	32888	32888	G65577	18q22.3	7		ZNF407	U
A050	60684	60684	G65578	05q23.1	7		PTD002	U
A051	220687	220687	G65579	12q24.31	7		NCOR2	O
A052	245287	116527	G65580	11q22.3	10		GRIA4	N
A060	103334	103334	G65639	05q11.1	8			R
A080	32634	32634	G65657	01p36.21	9			U
A081	33536	33536	G65658	22q12.3	8		TIMP3	U
A082	33654	-	G65659	10q23.31	19		TNFRSF6	O
A083	40518	40518	G65660	14q22.3	17			R
A090	40700	121688	G65667	15q15.3	8	FLJ35867		U
A092	33553	33553	G65669	13q13.3	9		DCAMKL1	R
A113	176061	435610	G65682	10p12.1	6		WAC	U
A133	40568	40568	G65700	15q13.1	7			U
A179	138944	138944	G65723	15q25.1	7			N

[a] This data was retrieved in June, 2003.

[b] R: retina-specific; N: neuronal; U: ubiquitous or expressed in several of the tissues tested; O: no expression in the cDNAs tested

n.a. : does not align to the current human genome sequence

Table 9 Characteristics and expression of retina EST-enriched clusters selected for expression analysis in UniGene Phase I

Lab ID	Original Hs. no.	Current Hs. no.[a]	GenBank Acc. No.	Chr. localization[a]	No. of retina ESTs	Ratio[c]	Reported gene or full-length sequence[a]	In intron of:[a]	*In-vitro* expression[b]
B1									
A104	8341	306676	G65674	20p13	1	0.33	FLJ14302		U
A105	20935	20935	G65675	01p31.2	1	0.33	DKFZp761D221		U
A107	24901	368477	G65677	06p2.3	1	0.33			U
A108	28487	-	G65678	12p11.22	1	0.50	FLJ10462 or LOC51290		U
A115	38705	38705	G65684	9p24.1	1	0.33		GASC1	U
A118	40528	40528	G65687	10q24.1	1	0.33		SORBS1	U
A119	40608	40608	G65688	11q25	1	0.50			U
A121	45119	227277	G65690	02p21	1	0.33	SIX3		R
A124	60563	60563	G65692	18q22.3	1	0.33	NETO1		R
A148	59844	59844	G65719	07q33	1	0.50			U
A150	47317	112921	G65711	15q21.3	1	0.50	DKFZp547H074		N
A185	14048	437920	G65604	11p15.4	1	0.50			N
A187	29553	29553	G65605	03q23	1	0.33		RASA2	U
A188	32489	32489	G65724	02q26.2	1	0.50			U
A189	60681	60681	G65606	08q13.3	1	0.50			N
A190	60772	60772	G65607	02q32.2	1	0.50			U
A191	60857	60857	G65608	22q13.31	1	0.33		CGI-51 or PARVB	O
A192	61119	61119	G65609	01q31.3	1	0.50	Probable CRB1 splice variant		N
A193	63356	63356	G65610	11q22.2	1	0.33			U
A194	66793	66793	G65611	06p24.2	1	0.33		MAK	R
A195	71023	71023	G65612	03q25.3	1	0.50			O
B2									
A054	31110	31110	G65636	Xq13.1	2	0.33		MGC34827	U
A065	129907	129907	G65644	01q24.1	2	0.50	C1orf32		N
A068	226925	226925	G65647	01q42.2	2	0.40	DKFZp686A202		U
A106	22979	98927	G65676	15q24.1	2	0.40	FLJ13993/ LMAN1L		N
A110	32677	32677	G65680	10q26.2	2	0.66		BCCIP and DDX32	U
A116	40183	40183	G65685	01q42.13	2	0.33			U
A122	49409	49409 & 64968	G65691	10q26.3	2	0.50			U
A123	61117	61117	G65727	10q23.31	2	0.66			U
A196	40515	303055	G65613	14q32.11	2	0.50		TTC8	N
A197	62528	404237	G65614	10p12.1	2	0.50		AB033043	U
A198	94327	94327	G65615	15q13.1	2	0.33		TJP1	O
A199	275335	- (H92006)	G65616	17q25.1	2	0.50			N
A205	61050	61050	G65621	02p13.2	2	0.40	AK093479		R
A207	125827	125827	G65623	03p26.31	2	0.33		NLGN1	R
A208	117926	117926	G65624	10q26.11	2	0.66		SLC18A2	U
A209	125070	440387	G65625	06q23.1	2	0.50	Probable EPB14L2 splice variant		R
A210	271692	271692	G65626	10q25.2	2	0.33	Probable PDCD4 splice variant		R
A211	271115	391367	G65634	16q22.1	2	0.40	Probablle SF3B3 splice variant		R
A213	240078	-	G65628	02q21.1	2	0.50	Probable CRYPTIC splice variant		N
A214	250664	250664	G65725	04p16.3	2	0.66	Probable RNF3 splice variant		R
A215	251827	302958	G65629	14q23.3	2	0.50		GPHN	U
B3									
A010	32840	32840	G65557	17q23.2	4	0.50		MSI2	R
A020	114256	114256	G65565	01p36.22	5	0.83		PEX14	O
A097	27186	192922	G65728	14q24.2	4	0.31	DPF3		U
A099	220536	220536	G65671	13q34	4	0.80		GAS6	U
A109	32452	32452	G65679	04q32.2	4	0.31	DKFZp566D234, KIAA1263		N
A111	32766	32766 & 433134	G65681	18q23	4	0.40			R
A117	40249	131886	G65686	12q24.31	3	0.75	BRI3BP		U
A120	40794	40794	G65689	15q24.1	3	0.38			O
A126	64616	64616	G65694	12p13.31	3	0.43	C12orf3		R
A142	60802	60802	G65706	07q21.1	5	0.83			U
A151	15260	443388	G65720	19q13.43	5	0.38			U
A181	117927	117927	G65600	06q21	4	0.50		REV3L	U
A182	269210	269210	G65601	11q23.3	4	0.50		KIAA0999	N
A183	250639	250639	G65602	15q14	3	0.75			R
A184	269208	269208	G65603	01p36.2	4	0.57		KIF1B	N
A200	40836	40836	G65617	11q22.3	4	0.57			U
A202	40838	40838	G65618	11q23.2	4	0.57		KIAA0999	U
A203	113872	144794	G65619	20p13	5	0.50	Probably STK35		R
A204	113876	113876	G65620	04p16.3	3	0.60		WHSC1	U
A212	233502	233502	G65627	02p21	4	0.67		MSH2	U
A218	32795	32795	G65632	07p15.2	5	0.71			O
B4									
A022	40808	40808	G65633	02p22.1	16	0.46	MGC33926		U
A023	250638	250638	G65566	12q24.33	47	0.63	MGC42193		U
A038	28411	28411	G65568	12q13.13	8	0.80	C12orf7		R
A039	61126	61126	G65569	09p11.2	6	0.75		BC006438	U
A040	227583	433652	G65570	Xp11.23	31	0.44	PPP1R3F		U
A041	33791	21413	G65726	20q13.12	17	0.94	SLC12A5		N
A044	32756	32756	G65572	04p16.1	7	0.87		FLJ31564	U
A048	59956	59956	G65576	03p26.1	8	0.89	GRM7		N
A053	29952	29952	G65635	Xp22.33	10	0.43			U
A055	32794	32794	G65715	07p21.1	13	0.93		Probably MGC42090	O
A058	25223	25223	G65637	01q31.2	14	0.47		Probably UCHL5	U
A059	40488	40488	G65638	17q11.22	10	0.83			R
A061	33827	33827	G65640	18q21.31	12	0.92		ATP8B1	O
A062	40400	44351	G65641	15q23	8	0.80		MYO9A	R
A063	40947	14202	G65642	17q11.22	10	0.91	MIRO-1		U
A064	63063	439217	G65643	01p21.3	9	0.90			U
A066	146060	146060	G65645	15q23	8	0.80		MYO9A	R
A067	221513	221513	G65646	06q21	8	0.80	AF520419		N
A069	28043	28043	G65648	04q12	8	0.67	Probably KDR		N
A070	32753	- (AA019309)	G65649	n.a.	9	0.90			R
A071	35493	35493	G65650	multiple	10	0.45	F379		R
A072	266470	266470	G65651	05q31.2	10	0.38	PCDHB4		U
A073	40905	40905	G65652	02q24.3	8	0.89			R
A075	59839	-	G65653	19q13.2	14	0.56			U
A076	107699	107699	G65654	06q13	11	0.61	B3GAT2 or SMAP1		U
A078	271783	271783	G65655	07q21.3	16	0.94			U
A079	247311	247311	G65656	02q33.1	10	0.31	FLJ33282		U
A084	62359	62359	G65661	10q21.1	12	0.60	BC036453 or KIAA1796		N
A098	60677	60677	G65670	13q32.1	6	0.67		MBNL2	U
A138	40507	40507	G65703	07q32.1	7	0.64		p100	O

[a] This data was retrieved in June, 2003.
[b] R: retina-specific; N: neuronal; U: ubiquitous or expressed in several of the tissues tested; O: no expression in the cDNAs tested
[c] Ratio = no. retina ESTs / total no. of ESTs in cluster
n.a. : does not align to the current human genome sequence

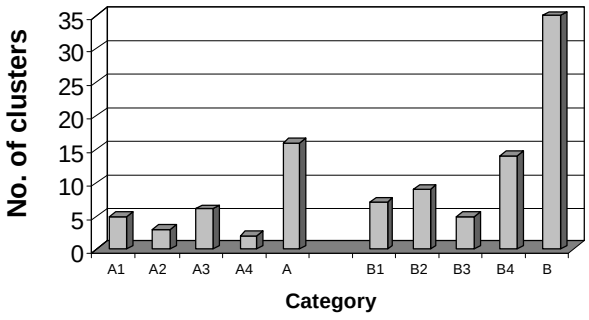
A repeated observation was that many of the cluster sequences (44%) are located in the intron of a known gene. Two genes may be encoded in the forward and reverse strand of a locus, but it is more probable that the observed phenomenon is due to retention of unspliced products in the cDNA library or to cloning of contaminating genomic DNA. The proportion of clusters mapping to introns was not significantly different for categories A and B (Table 10). The subcategory with the highest number of sequences localized in introns is A4. In contrast, only 24% of the B1 clusters map to introns of other genes.

Table 10 Summary of information for UniGene Phase I clusters

Category	No. of clusters	No. of genes that have been cloned[a] (%)	No. of clusters located in the intron of a known gene (%)
A1	22	5 (23)	9 (41)
A2	21	3 (14)	11 (52)
A3	22	6 (27)	8 (36)
A4	22	2 (9)	13 (64)
A_{total}	87	16 (18)	41 (47)
B1	21	7 (29)	5 (24)
B2	21	9 (43)	8 (38)
B3	21	5 (24)	9 (43)
B4	30	14 (47)	9 (30)
B_{total}	93	35 (34)	31 (33)
Total	**180**	**51 (28)**	**80 (44)**

Presently the genes of 67 clusters have been cloned and characterized. A significant difference is observed between the number of cloned genes from categories A and B. Whereas only 18% of the original clusters from category A are identified as cloned genes the percentage for category B is 34% (Fig. 7). From the latter, the subgroup with the highest cloning rate is B4, where 47% of the clusters have been identified as genes (Fig. 7).

Fig. 7 Number of cloned genes from clusters analyzed in UniGene Phase I
To date, the gene sequences of 51 of the 180 Hs. clusters are known and reported. The figure shows the exact number of cloned genes from each subcategory (A1-A4 and B1-B4) and also the total for each category (A and B). Significantly more genes have been cloned from the B category (34%) compared to category A (18%)

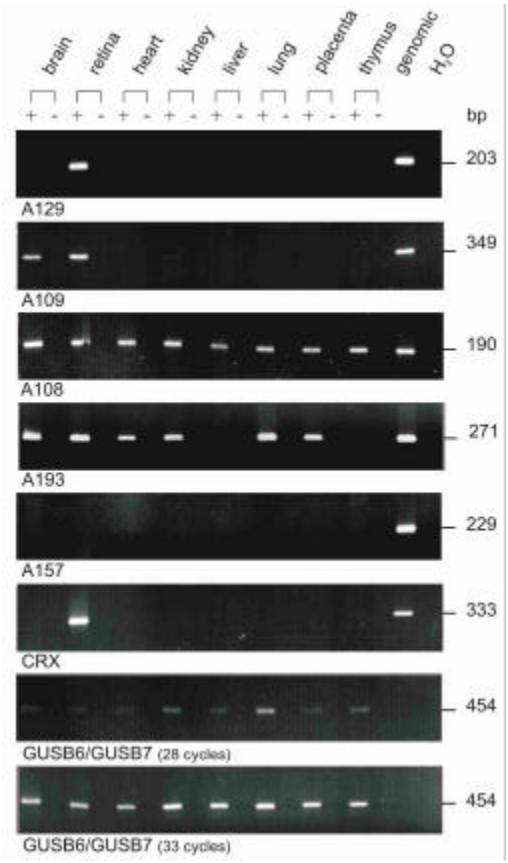


1.4.1. Expression analysis of UniGene clusters selected in phase I

From the 218 clusters for which primers were designed, 38 primer pairs could not be optimized and therefore expression analysis was done for only 180 genes (Table 11). The expression analysis of the selected clusters was accomplished by RT-PCR in a panel of eight DNAse-treated mRNAs. In addition to brain, retina, heart, kidney, liver, lung, placenta and thymus mRNAs, genomic DNA and water were

included as positive and negative controls, respectively. The quality and quantity of cDNA used for each tissue was estimated by amplification of the ß–glucuronidase (GUSB, LocusLink 2990) gene with primers GUSB3/GUSB5 and GUSB6/GUSB7 (Table 35) which amplify the 5'- and 3'-end of the gene, respectively. The amplification with GUSB3/GUSB5 also enabled us to evaluate the presence of any contaminating genomic DNA, since the product amplified from cDNA is 180 bp shorter than the genomic product. The applicability of the cDNA panel for the detection of retina-specific expression was confirmed by amplification of the retina-specific (Furukawa et al. 1997) cone rod homeobox protein gene (CRX) (Fig. 8) and the rhodopsin kinase gene (Zhao et al. 1998).

Fig. 8 Expression analyses of selected UniGene clusters
Representative examples of the various types of expression profiles found for the 180 clusters analyzed. Expression was demonstrated in retinal tissue only (A129), in retina and brain (A109), in all tissues tested (A108), in some but not all tissues tested (A193) and in genomic DNA only with no expression observed in all cDNA samples analyzed (A157). Retina-specific expression was confirmed with the photoreceptor-specific CRX as a positive control. RT-PCR amplification of the ubiquitously expressed GUSB (with GUSB6/GUSB7 primers) at 28 and 33 cycles facilitated normalization of the cDNA panel. RT-PCR reactions were performed in the presence (+) or absence (–) of reverse transcriptase. The size of the product is indicated on the right of the pictures.



Expression profiling of the 180 UniGene EST clusters identified 39 clusters whose sequence could only be amplified in retina cDNA (Table 11) and 32 clusters with transcription in the neural tissues (Table 11). Together these account for 40% of all clusters analyzed (Table 11). A total of 78 clusters were amplified in several or all tissues analyzed (Table 11) and for 31 clusters no RT-PCR products were obtained in any of the tissues tested, although PCR with genomic DNA amplified the expected fragment sizes (Table 11).

Comparison of the RT-PCR expression results of category A (retina EST-specific) and category B (retina EST-enriched) revealed no differences in the proportion of clusters exhibiting retina-specific (21 versus 23%) or neural expression (17 versus 18%) (Table 11).

Table 11 Summary of the expression profile of 180 UniGene clusters by category

Category	Ratio	No. retina ESTs	No. of clusters	No. of clusters expressed in			
				retina only (%)	neuronal tissues (%)	several tissues (%)	no expression (%)
A1	1	1	22	5 (23)	4 (18)	7 (32)	6 (27)
A2	1	2	21	4 (19)	6 (28)	5 (24)	6 (28)
A3	1	3-5	22	5 (23)	1 (5)	10 (45)	6 (27)
A4	1	=6	22	4 (18)	4 (18)	10 (45)	4 (18)
A_{total}			**87**	**18 (21)**	**15 (17)**	**32 (37)**	**22 (25)**
B1	<1 but = 0.3	1	21	3 (14)	4 (19)	12 (57)	2 (10)
B2	<1 but = 0.3	2	21	6 (28)	5 (24)	9 (43)	1 (5)
B3	<1 but = 0.3	3-5	21	5 (24)	3 (14)	10 (48)	3 (14)
B4	<1 but = 0.3	=6	30	7 (23)	5 (17)	15 (50)	3 (10)
B_{total}			**93**	**21 (23)**	**17 (18)**	**46 (49)**	**9 (10)**
Total			**180**	**39 (22)**	**32 (18)**	**78 (43)**	**31 (17)**

1.5. Phase II analysis of selected UniGene clusters

In the first stage of the UniGene analysis, a total of 342 Hs. clusters from a selection of 1241 retina-specific or enriched clusters were analyzed *in-silico*. From these, expression analysis was done for 180 clusters. Since the goal of the project was to establish a comprehensive list of retina-specific or abundant genes, we decided to proceed by selecting candidates from the remaining 899 Hs. clusters. As already indicated in the results of the first phase of the UniGene analysis, there was no clear and direct relationship between cluster composition and expression profile. Therefore, selection of clusters for expression profiling based on their composition was not pursued. From the experience of the first phase it was also clear that many clusters may contain artefactual sequences. Consideration of these factors, plus the bioinformatic advances and availability of the human genome resources enabled us to design a new strategy in order to proceed as efficiently as possible. The goal for this stage was to analyze all Hs. clusters with two or more retina ESTs in a proportion greater than 30% of the total EST content. Therefore 293 Hs. clusters were included in the Phase II analysis.

The strategy for Phase II relied more heavily on bioinformatical analyses compared to Phase I. The goal was to be able to predict as accurately as possible the expression and veracity of a cluster sequence based on *in-silico* information. Those considered eligible for further analysis were chosen for *in-vitro* expression studies. This work was greatly facilitated by the newly released Genome Browser[43] from the University of California Santa Cruz (Kent and Haussler 2001, Kent et al. 2002). The browser provides a rapid and reliable display of any requested region of the genome, together with dozens of aligned annotation tracks (known genes, predicted genes, ESTs, mRNAs, homology to the mouse genome, assembly gaps and coverage, chromosomal bands, etc.). The annotation tracks are found beneath genome coordinate positions, allowing rapid visual correlation of different types of information (Fig. 9). This allows the user to retrieve a multitude of facts related to the existence and organization of a gene.

[43] http://genome.ucsc.edu/cgi-bin/hgBlat?command=start

The analysis was started by revising the 293 Hs. clusters on UniGene build #145 (December 2001). No further analysis was possible for 57 of the clusters whose EST sequences were not included in any of the current UniGene clusters. In case that an EST had been reorganized in a new cluster, the thread was followed and the new Hs. cluster was analyzed even though some did not comply anymore with the 30% retina EST content cut-off established at the beginning of the project. In short, 105 clusters were still retina-specific in their composition, and 78 were retina-enriched (clusters where at least 30% of the ESTs were derived from retina). At the time of re-evaluation, 53 clusters had an enrichment of retina ESTs lower than 30%.

For each of the 236 clusters analyzed *in-silico* a wealth of information was recorded in the database. The recompilation began by noting the 5'- and 3'-THC ID analogue to the Hs. cluster and the existence of repeats in these sequences. It was set forth by comparing the sequences to the human genome version (November 2001) and recording overlaps to known genes or gene predictions, mapping to the intron of a gene, presence of splice sites, similarities with sequences of other species, etc. These observations serve as proof that the sequences belong to transcribed sequences and were relevant for the decision-making process.

Candidates for expression analysis were chosen after categorization in nine classes taking into account the surveyed information in combination with the picture depicting the genomic contig around the sequence. Eighteen sequences aligned to predicted genes, did not contain repeats, had high homology to non-human mRNAs, and revealed accurate splicing behaviour. Those transcripts were considered to have excellent matches to the defined parameters and were included in class 1 (Fig. 9). A greater number of sequences (32 clusters) complied in part with the above mentioned parameters and were grouped in class 2 (Fig. 9). The majority of the clusters (50) were included in class 3 which included clusters fulfilling only a minimum of the criteria (Fig. 9). The sequences of twenty-six clusters consisted primarily of repeats and were included in class 4. If the cluster sequences were identical in part to a full-length clone (e.g. KIAA, DKFZ, etc.), as was the case for 21 clusters, they were grouped in the fifth class. Class six was reserved for ten sequences that did not overlap but were very near to full-length clones. Even though the human genome sequence draft was already published, four of the cluster sequences did not align to the human draft and were grouped in class 7. Finally, since more than a year had passed since the primary UniGene survey 18 of the clusters were labeled as known genes (class 8) and positional analysis of the remaining 27 Hs. clusters revealed that they probably derive from known genes (class 9).

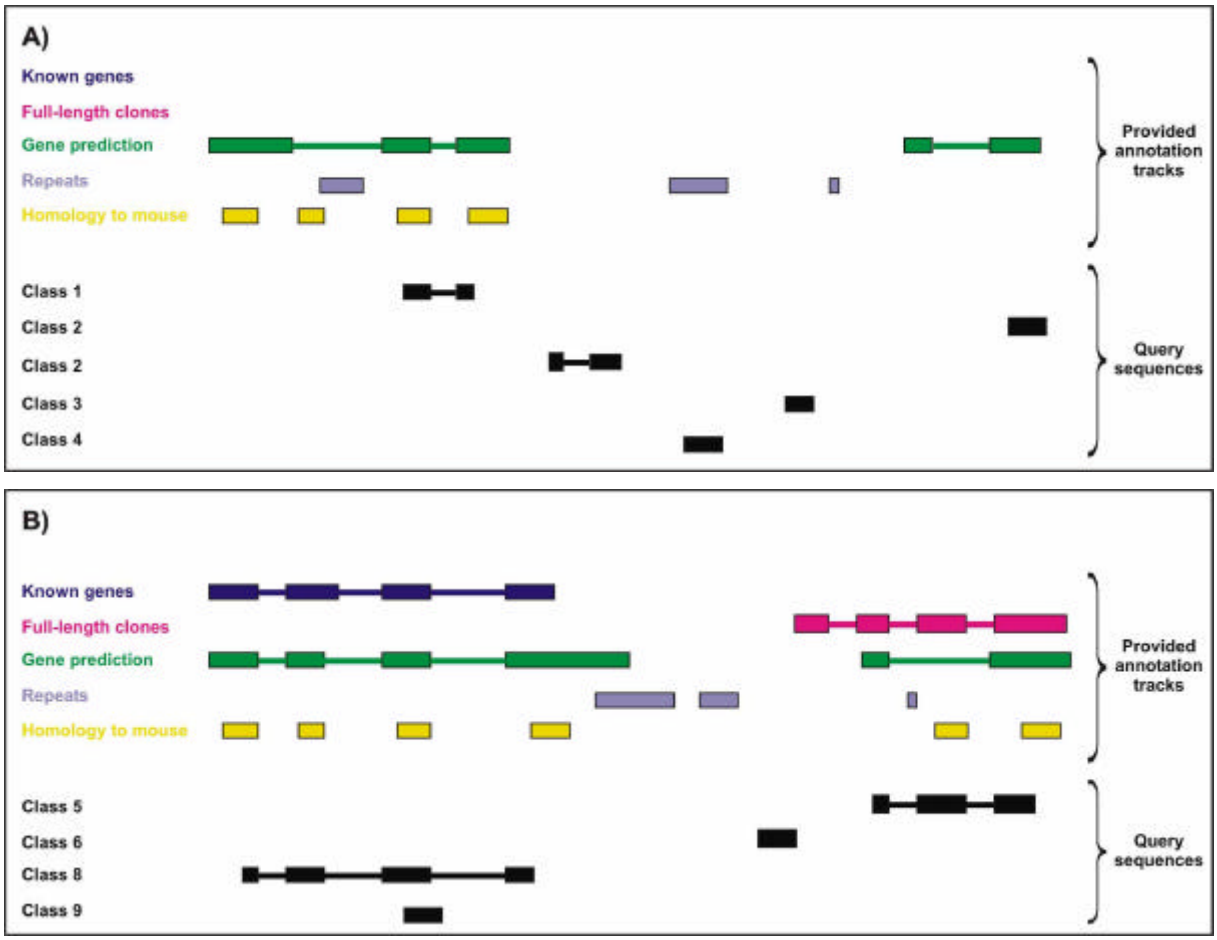


Fig. 9 Schematic examples of the outputs used to catalogue Phase II clusters in classes
The colored annotations represent a selection of the tracks provided by the Genome Browser USCS. Examples of query sequences are shown in black. A) All sequences which do not overlap or map near to known genes or full-length clones were grouped in classes 1 to 4 following the criteria explained in section 1.5. In order to illustrate the various possibilities, two examples of clusters from class 2 are shown. B) All sequences overlapping to full-length clones were grouped in category 5, those mapping near to such sequences were grouped in category 6. No example of category 7 clusters is shown because these are the clusters that did not align to genomic sequence. Class 8 contains clusters that derive from known genes, whereas class 9 contains those that probably belong to known genes but contain sequences that are not 100% similar to the known gene sequence.

Based on this *in-silico* analyses, clusters for which there was sufficient evidence that it derived from a real transcript but had not yet been identified as a gene were chosen for expression analysis. Primers for expression analysis were designed for 73 Hs. clusters from classes 1, 2, 5, and 6. From the 73 clusters, 24 contained only retina ESTs, whereas 25 of the clusters had a retina EST proportion greater than 30%, and 24 had at the time a proportion of retina ESTs smaller than 30% (Table 12).

Table 12 Summary of clusters selected for expression analysis in UniGene Phase II

Lab ID	Original Hs. no.	Current Hs. no.[a] (EST acc. no.)	Chr. localization[a]	No. of retina ESTs	Ratio[c]	Reported gene or full-length sequence[a]	In intron of:[a]	Homology to mouse	Class	*In-vitro* expression[b]
B001	101672	101672	04p16.1	4	0.22	FLJ31564		y	1	N
B002	110293	110293	10p15.1	4	0.29			n	1	U
B003	114263	114263	02q34	4	0.80	FLJ33496		y	1	A
B004	161043	123648	07q34	2	0.29	FLJ25778		y	1	U
B005	171895	171895	18p11.32	7	0.47			y	1	U
B006	188962	172792	02q32.1	2	0.25	KIAA1946		y	1	N
B007	191637	191637	12q23.3	3	0.38			y	1	U
B008	193815	193815	03q12.3	4	0.80			y	1	U
B010	216717	216717	06q14.1	6	0.44			n	1	U
B011	233359	- (H40707)	22q11.22	7	0.50		BC022822	y	1	A
B012	268793	- (R93853)	12q13.2	3	1.00			y	1	O
B013	284998	284998	10p13	2	1.00			n	1	U
B014	310649	- (AA020898)	01p22	2	0.25			y	1	A
B015	194617	196582	14q22.1	1	0.23	C14orf29		y	1	N
B016	14235	14235	01p21.1	3	0.05	FLJ25070		y	5	U
B017	23744	23744	05q34	7	0.03	PANK3		y	5	U
B018	33619	33619	05q23.2	3	0.32			y	5	U
B019	39850	39850	20q13.33	8	0.02	URKL1		y	5	U
B020	60687	294151	17q25.1	3	0.07			y	5	U
B021	69559	69559	01q24.3	7	0.01	XTP2		y	5	U
B022	72660	72660	17q25.1	4	0.02	PTDSR		y	5	U
B023	118554	118554	08q13.3	3	0.16	CGI-83		y	5	U
B024	136315	136315	11q13.5	19	0.08	AQP11		y	5	U
B025	144609	144609	03q27.2	2	0.05	MGC15397		y	5	U
B026	173958	173958	09q34.3	2	0.19	PRO2405		y	5	U
B027	181173	181173	02q35	4	0.32	MGC10771		y	5	U
B028	243901	243901	08q22.1	9	0.02			y	5	U
B029	293678	293678	19q13.33	9	0.06	TCBAP0758		y	5	A
B030	294151	294151	17q25.1	9	0.16	KIAA1917		y	5	N
B031	331552	284232	01p36.31	16	1.00	KIAA0720		y	5	U
B032	40489	40489	11p15.1	6	0.57			y	2	R
B033	40505	40505	16q12.1	4	1.00			y	2	U
B034	40840	50150	11q24.1	8	0.75		AK091713	y	2	U
B035	60556	- (H85926)	03p12.1	6	1.00		LOC253559	y	2	R
B036	60768	425096	05p15.31	9	1.00			y	2	R
B037	60836	60836	14q24.3	2	1.00			y	2	N
B038	60943	60943	15q26.6	4	1.00			y	2	U
B039	164577	- (AA017283)	15q26.2	2	0.36			n	2	U
B040	107680	107680	14q11.2	4	1.00		ZNF409	y	2	U
B041	114261	114261	06q22.1	4	1.00			y	2	O
B042	169375	32793	02q32.3	4	0.18	FLJ31108		y	2	U
B043	172856	172856	01p36.2	2	0.44		AK092289	y	2	R
B044	185826	185826	13q33.3	4	1.00			n	2	U
B045	269211	269211	04p34.3	3	1.00			y	2	O
B046	185812	373854	08q21.2	4	0.33	GOR		y	2	O
B047	188687	188687	01q25.3	2	1.00	Partly RGL		n	2	U
B048	188948	188948	17q25.3	4	1.00		raptor	y	2	O
B049	189992	- (AA169184)	13q33.3	3	1.00			y	2	O
B050	190830	- (AA058913)	22q13.1	3	0.30			y	2	U
B051	202977	- (AA218628)	06p25.2	3	0.40			y	2	U
B052	211191	148845	01p34.2	2	0.40			y	2	R
B053	220688	312306	06q25.2	2	1.00			y	2	U
B054	221147	221147	01q42.3	4	1.00			y	2	A
B055	221924	221924	12p13.31	4	1.00		SLC2A3	n	2	U
B056	221934	221934	09p13.3	2	1.00			y	2	U
B057	237689	237689	08q24.22	2	0.32		D63477	y	2	U
B058	240934	- (H87838)	05q23.1	6	1.00			y	2	U
B059	271617	- (AA018527)	07q22.1	2	1.00			y	2	O
B060	336636	97296	08q24.3	3	0.30		AB020677	y	2	N
B061	3542	3542	07p21.3	3	0.04	FLJ11273		y	6	U
B062, A105a	32759	32759	01p31.2	7	1.00		DKFZp761D221	y	6	U
B063	33271	- (BM687107)	17q25.3	7	0.67		KIAA1453	y	2	U
B064	59821	59821	05p13.2	4	1.00	AK094895		y	6	U
B065	124154	124154	15q25.1	9	0.30		BC020256	y	6	U
B066	146243	279635 and 155182	14q24.3	3	0.33	KIAA1036		y	6	O
B067	151010	- (AA258656)	12q13.2	2	0.19	CIP29 UTR[d]		y	6	U
B068	174918	174918	02q14.1	3	0.50		FLJ10996	y	5	U
B069	214043	214043	06p23	2	0.29	FLJ20958 UTR[d]		y	6	U
B070	245292	245292	17p11.2	2	1.00		HCMOGT-1	y	6	U
B071	302750	- (AA056306)	12q13.12	5	0.37		LOC91012	y	6	O
B072	323230	323230	16p13.3	7	0.67	Partially RFWD1		y	6	O
B073	137514	148845	01p34.2	4	0.25			y	1	U
B074	145592	145592	07p22.3	1	0.40	KIAA1908		y	1	N

[a] This data was retrieved in June, 2003. The hyphen (-) indicates that no cluster contains the sequence anymore. In such cases, one EST which used to be in the original cluster is listed between parentheses

[b] R: retina-specific; N: neuronal; U: ubiquitous or expressed some of the tested tissues; O: no expression in the tested cDNAs

[c] Ratio = no. retina ESTs / total no. of ESTs in cluster

[d] The cluster probably derives from the listed gene

1.5.1. Expression analysis of UniGene clusters selected in phase II

For the RT-PCR expression analysis 73 primer pairs were designed with the Primer3 design software[44] and OLIGO version 2.0 tool (Rychlik and Rhoads 1989). If possible, the primers were chosen so that the product amplified from cDNA would differ from the genomic amplicon. In order to facilitate distinction between both phases of the UniGene project, the clusters and primers were identified with a lab ID based on the letter B followed by a number (e.g. B001). The optimization of 68 primer pairs was successful, but it was not possible to optimize PCR conditions for B003, B011, B014, B029, and B054. For each cluster the expression was first established with a cDNA panel containing seven different tissues (brain, heart, lung, placenta, retina, thymus, and RPE), genomic DNA, and a negative control (water). For 17 clusters (B001, B006, B015, B019, B020, B023, B030, B032, B035, B036, B037, B040, B042, B043, B052, B060, and B067) with an interesting expression a more comprehensive expression profiling round was done in a panel containing 20-tissues (Fig. 10).

Even though a careful pre-selection had been done, only five of the 68 clusters (7%) are expressed exclusively in retina; neuronal expression was found for seven (10%) of the genes. Forty-six genes (73%) are ubiquitously expressed and 10 primer pairs amplify a product only from the genomic DNA template.

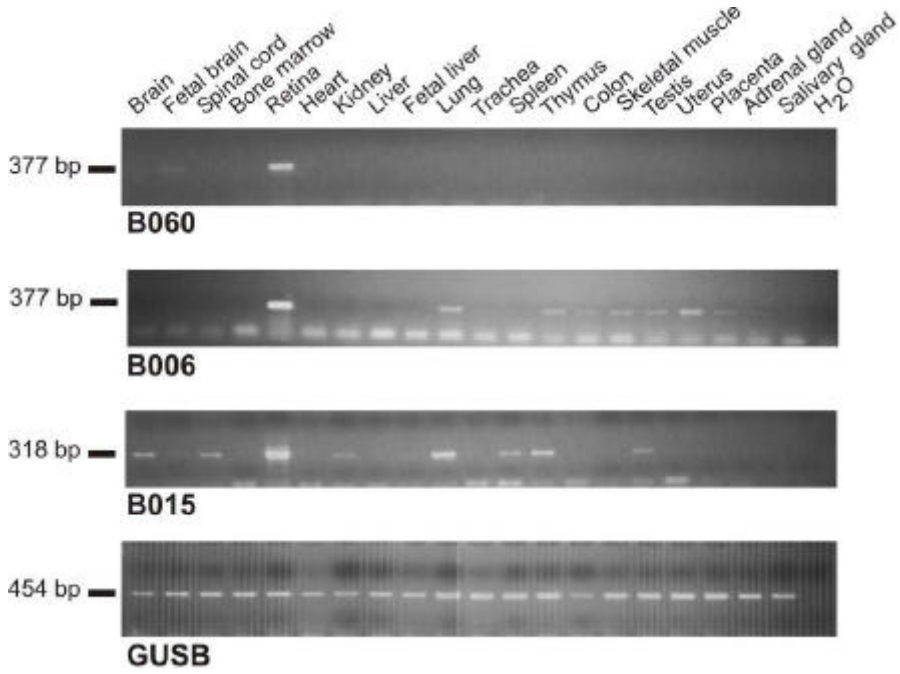


Fig. 10 Representative examples of expression profiles found in Phase II analysis.
B060 is highly expressed in retina and weakly in fetal brain. The B006 transcript is expressed predominantly in retina, but there is also some expression in other tissues. Something similar, but coupled with neuronal expression is seen for B015. The quality of the 20-tissue panel was calibrated by RT-PCR amplification of the ubiquitously expressed GUSB gene (with GUSB3/-GUSB5 primers)

[44] http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi

2. Analysis and data-mining of the retina suppression subtracted hybridization retina cDNA library (retSSH)

2.1. Analysis of library complexity

The retina suppression subtracted hybridization cDNA library was generated by Andrea Gehrig using the SMART cDNA synthesis and long-distance PCR (LD-PCR) amplification strategy (Barnes, 1994). The enrichment of transcripts expressed preferentially in retina was achieved by using retina cDNA as tester and kidney and liver cDNAs as drivers.

To characterize the library and identify novel retina-enriched genes, 1093 randomly picked clones were sequenced partly in-home and also by a commercial supplier. The sequences were compared with data stored in the non-redundant nucleotide, EST and human genomic sequence collections using BLAST algorithms[45]. All sequences derived from the mitochondrial genome or containing vector sequence only were discarded, thus the number of clones which were further analyzed in all future steps was 1080. As expected, comparison of the BLAST results indicated redundancy of the cDNA sequences. Consequently, all clones derived from the same gene were grouped in a cluster. The clustering process, done in March 2003, revealed that 931 out of 1080 analyzed clones represent 321 known genes. The clones with no homology to any of the genes found in the LocusLink[46] collection were catalogued as unknown genes. Clustering of these transcripts was based on a comparison of the sequences with each other as well as positional analysis to identify sequences which derived from the same genomic region. Based on the gathered information, 92 clusters were assembled from the 149 clone sequences. In total, 76% of the unknown gene sequences could not be assembled in clusters and are hereafter referred to as singletons (Table 13). While it is known that singletons may arise from artefact sequences present in the library such as human genomic DNA contamination or unprocessed mRNA, this does not appear to be a major factor in our collection of unknown transcripts. This is supported by the fact that the percentage of singletons from the known and unknown gene categories is not significantly different (Table 13). The same applies for clusters containing two sequences. The overall redundancy of this library is approximately 2.6 clones per transcript.

Table 13 Clustering statistics of the clones sequenced from the retSSH library

No. of clones per cluster	Known genes			Unknown genes		
	No. of clusters		No. of clones	No. of clusters		No. of clones
1	218	(68%)	218	70	(76%)	70
2	43	(13%)	86	12	(13%)	24
3	22	(7%)	66	5	(5%)	15
4	8	(2%)	32	0	(0%)	0
5	6	(2%)	30	2	(2%)	10
= 6	24	(7%)	499	3	(3%)	30
Total no. of clusters	321		931	92		149

[45] http://www.ncbi.nlm.nih.gov/BLAST/
[46] http://www.ncbi.nlm.nih.gov/LocusLink/

The mRNA species detected more than 10 times account for 43% of the sequences analyzed but represent only 5% of the total number of unique genes identified in this study. This discrepancy is explained by the fact that prevalent and intermediate frequency mRNAs comprise as much as 50-60% of the total mRNA (Bonaldo et al. 1996). As expected in a subtracted library, 17 of the 19 genes whose clusters have 10 or more clones are neuronal-specific according to the literature and/or Gene Expression Atlas[47] (Table 14). The cluster with the highest number of clones (75) represents the guanine nucleotide binding protein, gamma transducing activity polypeptide 1 (GNGT1); it is followed by S-antigen (SAG), and phosphodiesterase 6A, cGMP-specific, rod, alpha (PDE6A). Another indicator that the subtraction was efficient is the fact none of the clones contained sequences from housekeeping genes and only one clone contained a ribosomal protein sequence (RPS27A).

Table 14 Most frequently found genes in the retSSH library

Gene symbol	Gene name	No. of clones	Neuronal-specific[b]
GNGT1	Guanine nucleotide binding protein, gamma transducing activity polypeptide 1	75	yes
SAG	S-antigen (arrestin)	56	yes
PDE6A	Phosphodiesterase 6A, cGMP-specific, rod, alpha	42	yes
CLUL1	Clusterin-like 1 (retinal)	41	yes
PDC	Phosducin	41	yes
RHO	Rhodopsin	35	yes
PDE6H	Phosphodiesterase 6H, cGMP-specific, cone, gamma	22	yes
IMPG2	Interphotoreceptor matrix proteoglycan 2	19	yes
TPD52	Tumor protein D52	15	no
GLUL	Glutamate-ammonia ligase (glutamine synthase)	15	no
GPM6A	Glycoprotein M6A	14	yes
uL33[a]	-	14	yes
GUCA1C	Guanylate cyclase activator 1C	13	yes
SLC1A3	Solute carrier family 1, member 3	12	yes
IMPG1	Interphotoreceptor matrix proteoglycan 1	12	yes
MAP2	Microtubule-associated protein 2	12	yes
CNGA1	Cyclic nucleotide gated channel alpha 1	10	yes
uL34 [a]	-	10	yes
HMGN1	High-mobility group nucleosome binding domain 1	10	no

[a] Lab ID gene symbol.
[b] Expression reported in the literature or Gene Expression Atlas (http://expression.gnf.org/cgi-bin/index.cgi)
Gray shade indicates ubiquitously expressed genes.

[47] http://expression.gnf.org/cgi-bin/index.cgi

2.2. Expression profiling of the retSSH library

2.2.1. Classification of the retSSH clusters

To select candidates for expression profiling the 413 clusters assembled from the 1080 sequences were analyzed *in-silico* and divided in four categories by Jelena Stojic in February, 2002. The first category included all the clusters which contained the sequence of an already known gene (230 clusters). The 76 clusters that had the sequence of an already reported full-length mRNA (e.g. KIAA, DKFZ, etc.) were categorized in the second group. The third category included 16 clusters whose sequences spliced at least once. Category 4 included the remaining 91 clusters which were not similar to any mRNA and did not splice.

2.2.2. Expression analysis of selected retSSH clusters

The expression profiling of the 94 clusters included in categories 2 and 3 was done by Jelena Stojic. Each of the investigated clusters received a lab ID consisting of the letter L followed by a number (e.g. L39) in order to distinguish them from the UniGene clusters. The expression was investigated by RT-PCR in a panel consisting of nine cDNAs (brain, heart, lung, retina, RPE, placenta, thymus, uterus, and fibroblast) plus a genomic and water control. Fifteen transcripts expressed exclusively in retina were identified. Another 10 genes are expressed in retina at higher levels than the remaining tissues and an additional 11 present neuronal expression (personal communication).

2.3. Evaluation of library completeness and enrichment

2.3.1. Assessment of the amount of sequenced clones

To evaluate whether the sequenced clones are representative of the library complexity, the 1080 clones were randomly sorted and packed in 27 groups of 40 clones each. For each ‘plate’ the number of unique and redundant genes was computed (Table 15). By assigning an ID to the unknown gene clusters, they could also be included in the analysis even though the genes have not been cloned yet. The number of unique genes per 40-clone plate ranged from 28 to 38. Even though the number of new genes decreases as more clones are sequenced, the reduction is not statistically significant and many new genes can be found with each 40 clones that are sequenced (Fig. 11).

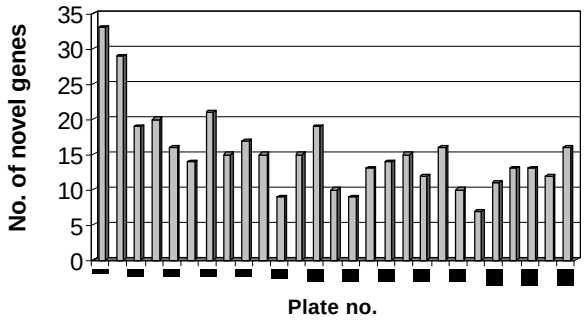


Fig. 11 Number of novel genes per plate
As expected, the number of novel genes per plate decreases after the first plates but the number stays fairly constant for all the other plates.

Table 15 Analysis of redundancy and new genes found in random plates

Plate no.	Cumulative no. of clones	Cumulative no. of genes	Genes per plate	Novel genes per plate
1	40	33	33	33
2	80	62	33	29
3	120	81	28	19
4	160	101	33	20
5	200	117	29	16
6	240	131	28	14
7	280	152	33	21
8	320	167	31	15
9	360	184	31	17
10	400	199	29	15
11	440	208	29	9
12	480	223	33	15
13	520	242	38	19
14	560	252	30	10
15	600	261	32	9
16	640	274	34	13
17	680	288	34	14
18	720	303	31	15
19	760	315	33	12
20	800	331	36	16
21	840	341	34	10
22	880	348	29	7
23	920	359	33	11
24	960	372	33	13
25	1000	385	33	13
26	1040	397	31	12
27	1080	413	32	16

2.3.2. Evaluation of the degree of subtraction

In a study published in 2001, Bortoluzzi et al. presented a list of ribosomal proteins and compared their *in-silico* expression in six different tissues. For our study, the list was slightly modified and included the 84 genes for which a gene symbol could be tracked. On the basis of the *in-silico* analysis Bortoluzzi et al. predict them to be expressed in many tissues, although at different levels. Hence, if a library has been satisfactorily subtracted there should be no copies of these ribosomal genes left. With this in mind the proportion of ribosomal genes found in the library was inspected. Since only one (RPS27A) of 84 ribosomal genes was found, we propose that the optimal subtraction of the retSSH library has been reached.

2.3.3. Evaluation of the library by comparison with known retinal pathways

The highly-specialized functions of the retina are due to a number of specific pathways involved in the conversion of light signals into electrical impulses which are then transmitted to the optical centres of the brain. Essential pathways for these functions are for example the vitamin A cycle, the phototransduction cascade, and the synaptic transmission of signals. A list of genes belonging to each of the pathways was compiled from published literature. The retSSH gene collection was then compared with these lists of pathway-associated genes to assess the quality and representation of the library.

2.3.3.1. Comparison with the phototransduction cascade and vitamin A cycle pathways

The partial list of genes expressed in retina and involved in phototransduction and the vitamin A cycle includes 53 genes. Nineteen (36%) of these genes were found in the retSSH library and are indicated by grey shading in the following list of genes involved in these pathways (Table 16).

Table 16 Genes associated with phototransduction and the vitamin A cycle

ABCA4	CNGA3	**GNGT1**	**GUCY2D**	**PDC**	**PDE6H**	**RBP3**	**RGR**	**RLBP1**
ALDH1A1	**CNGB1**	GNGT2	GUCY2F	**PDE6A**	PDEA2	RBP4	RGS11	**SAG**
ALDH1A2	CNGB3	GPRK7	NCALD	**PDE6B**	RARA	**RCV1**	**RGS16**	SLC24A1
ARR3	**CRABP1**	**GUCA1A**	**OPN1LW**	PDE6C	RARB	**RDH12**	RGS9	SLC25A18
CACNA1F	**GNAT1**	**GUCA1B**	**OPN1MW**	PDE6D	RARG	RDH5	**RHO**	**SLC25A22**
CNGA1	GNAT2	GUCA1C	OPN1SW	**PDE6G**	RBP1	**RDH8**	RHOK	

Shaded genes were found in the retSSH library
Bold genes were found in the retNEIBank library

2.3.3.2. Comparison with genes involved in synaptic transmission of neuronal signals

A list of 194 genes involved in synaptic transmission was compiled by searching the literature (Table 17). Although it includes genes which are expressed in synapses, in many cases direct evidence of retinal expression is missing. Many of the genes are part of large gene families and it is not readily possible to determine exactly which members of the family are involved in the synaptic pathway. In contrast to the phototransduction and vitamin A pathway, the proportion of genes from this list which has been found in the library was much smaller. Only 4% (8 from 194) of the genes were found in the retSSH library. The nine genes are NSF, SLC1A2, SLC1A3, SNAP25, STX12, STXBP1, SV2B, and SYT1 (Table 17, shaded).

Table 17 Genes involved in synpatic transmission of neuronal signals

ABAT	CLCN3	EFNB3	GABRA4	**GOT1**	GRM6	**NSF**	**SLC1A3**	**STXBP6**
ABLIM1	CLCN4	EPHA1	GABRA5	GOT2	GRM7	**OAT**	**SLC1A7**	SV2A
ADCY1	CLCN5	EPHA2	GABRA6	GPHN	GRM8	PCLO	SLC4A3	SV2B
ADCY2	CLTA	EPHA3	GABRB1	GPR51	GUCY1A2	PRKCA	SLC6A1	SYN1
ADCY3	**CLTB**	EPHA4	GABRB2	GPT	GUCY1A3	PRKCABP	SLC6A11	SYNGAP1
ADCY4	CORTBP2	EPHA4	GABRB3	GRIA1	GUCY1B2	**RAB3A**	SLC6A13	**SYNGR1**
ADCY9	CPLX1	EPHA5	GABRD	GRIA2	GUCY1B3	RAB3	SLC6A5	SYNGR2
AGRN	CPLX2	EPHA6	GABRE	GRIA3	GUCY2E	RIMS1	SLC6A9	SYNGR3
APBA1	CX36	EPHA7	GABRG1	GRIA4	HOMER1	RIMS2	SNAP23	SYNPR
APBA2	DLG1	EPHA8	GABRG2	GRIK2	HOMER2	RPH3A	SNAP25	**SYP**
ATP2B1	DLG2	EPHB1	GABRG3	GRIK3	HOMER3	RPH3AL	SNAP25	**SYT1**
ATP2B2	DLG3	**EPHB2**	GABRP	**GRIK5**	HOMER4	S100B	SNAP29	**UNC119**
ATP2B3	**DLG4**	EPHB3	GABRQ	**GRIN1**	**HPCA**	SCAM-1	**SNAPAP**	UNC13
ATP2B4	DLGAP1	EPHB4	GABRR1	GRIN2A	**HPCAL1**	SDK1	SSTR1	UTRN
BCAT1	DRD1	EPHB5	GABRR2	GRIN2B	HPCAL4	SDK2	SSTR2	VAMP1
BCAT2	EFNA1	EPHB6	**GDI1**	GRIN2C	KIF3A	SHANK2	SSTR4	**VAMP2**
BSN	EFNA2	FREQ	GJB2	GRIP1	LIN7A	**SHMT2**	STX12	VAMP5
CABP1	EFNA3	GABBR1	GLRA1	**GRIPAP1**	LIN7C	SLC12A5	STX1B1	VAMP8
CABP5	EFNA4	GABRA1	GLRA1	GRM1	**NAPA**	SLC17A6	**STX1B2**	
CASK	EFNA5	GABRA2	GLRA2	GRM2	NLGN1	**SLC17A7**	STX3A	
CLCN1	EFNB1	GABRA3	GLRA3	GRM4	NLGN2	SLC1A1	**STXBP1**	
CLCN2	EFNB2	GABRA3	GLRB	GRM5	NLGN3	SLC1A2	STXBP5	

Shaded genes were found in the retSSH library
Bold genes were found in the retNEIBank library

2.3.4. Estimation of the number of genes found in the library

The retSSH library was generated for two reasons. The major motivation was to identify novel retina-abundant genes, while a second important reason was to estimate the number of genes preferentially expressed in the retina.

Determination of the number of biologically significant genes in a particular cell or tissue is a challenging biological problem (Bishop et al. 1974, Velculescu et al. 1999, Stern et al. 2003). Even though many researchers have attempted to develop a model for the underlying gene expression level probability function (GELPF) that could be used to determine the statistical distribution of the number of genes expressed in a cell, there is still no satisfying approach (Kuznetsov et al. 2002). A difficult issue in the approximation is the existence of genes expressed at very low levels in the cell. Therefore the only way to estimate gene frequency and distribution is by relying on experimental evidence and attempting to extrapolate the observations.

Theoretically the retSSH library should contain at least one copy of each gene expressed at higher levels in retina than liver or kidney. Therefore, if the library were to be sequenced to exhaustion, it would be possible to estimate the number of genes expressed in retina at higher levels than other tissues. Although only a fraction of the retSSH library has been sequenced until now, we evaluated a statistical approach to predict the number of genes preferentially expressed in retina tissue.

To assess if the number of sequenced clones are sufficient to estimate the number of unique retina-abundant transcripts, the 27 random ‘plates’ were statistically analyzed. In a first step the number of unique genes discovered by sequencing the given number of clones was plotted (Fig. 12A). Since the slope of the curve is still ascending and has not reached a plateau it is not possible to estimate the number of genes preferentially expressed in retina. In order to calculate how many clones it would be necessary to sequence in order to find only one novel gene per 40 sequenced clones, the number of new genes per plate was plotted against the number of sequenced clones (Fig. 12B). Based on the confidence interval of the linear regression, between 10,000 and 300,000 clones would have to be sequenced before only one novel gene would be identified 40 sequenced clones.

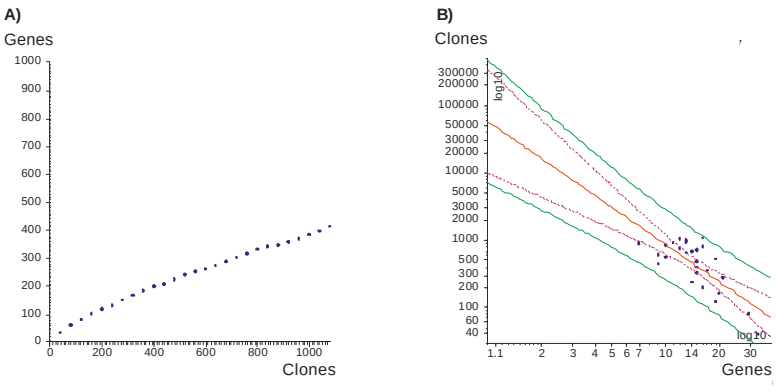


Fig. 12 Statistical analysis of the retSSH library

A) The cumulative number of unique genes discovered after sequencing 27 ‘plates’ of 40 clones each has been plotted.

B) The number of novel unique genes found in each of the plates is shown as a function of the total number of sequenced clones. Using the data points a regression line was calculated (red line). The pink line shows the 95% confidence interval of the regression line. The outer green line shows the 95% confidence interval of the individual points.

From the above stated it follows that the number of sequenced clones is not sufficient to elaborate a statistically solid model that would predict how many clones would need to be sequenced until only one novel gene is identified after sequencing 40 clones. Similarly, it is not possible at present to extrapolate the results to predict the number of retina-preferential genes.

3. Analysis of the NEIBank retina cDNA library (retNEIBank)

The quality of the retSSH library was assessed by comparison to another collection which served as reference. The only large and suitable collection of retina sequences found in a well organized database is provided by the NEIBank project. The retNEIBank collection[48] originated from the ongoing sequencing of an unamplified retina cDNA library (Wistow et al. 2002) and as of February 2003, a total of 2701 clusters assembled from 4913 sequences had been reported.

3.1. General characteristics of the retNEIBank library

The 4913 sequenced clones are presented in the database as clusters, with links to each of the sequences and many annotations (e.g. number of clones in cluster, gene annotations, etc.). All 2701 retNEIBank clusters were reanalyzed in order to update gene symbols and assign LocusID identifiers to each to facilitate future comparisons with the retSSH library. During this process we identified a number of genes for which more than one cluster had been created. As an example, clusters number 1037 and 2133 both contain sequences from the bridging integrator 1 gene (BIN1) and have the same UniGene cluster as reference. In such cases the clusters were merged and thereafter were considered as only one cluster. For this reason, our revised NEIBank database contained only 2615 clusters. From the 2615 unique transcripts 2114 correspond to known genes and 501 were yet unidentified genes. In total, 75% of the clusters are singletons. An analysis of the largest clusters was done applying the same criteria applied for selection of the largest clusters of the retSSH library. In total 40 clusters of the NEIBank library contain 43% of the total clones (Table 18).

Table 18 List of most abundant transcripts from the retNEIBank library

Gene symbol	No. of clones	Housekeeping*	Gene symbol	No. of clones	Housekeeping*	Gene symbol	No. of clones	Housekeeping*
RHO	138	no	PKM2	22	no	PSAP	16	yes
GPX3	72	yes	ROM1	22	no	NRL	16	no
PDE6G	60	no	ALDOA	21	yes	UBC	16	yes
EEF1A1	58	yes	RPL3	21	yes	HSPCA	15	yes
CLU	53	yes	TUBB2	21	no	RPS2	15	yes
GAPD	51	yes	VIM	21	yes	RPL4	15	yes
FTH1	50	yes	EEF1G	19	yes	PTGDS	14	no
TF	47	no	SPP1	19	no	ENO2	14	no
CKB	44	no	ACTG1	18	yes	RPL13	14	yes
SAG	33	no	RPL13A	17	yes	TPI1	13	no
RPS3A	33	yes	ATP1A3	17	no	GLTSCR2	13	no
RCV1	28	no	ALDOC	17	no	AIPL1	13	no
GLUL	26	yes	UNC119	17	no			
GNAT1	25	no	RPLP0	16	yes			

Bold names indicate genes that are also found in the retSSH gene list of abundant genes
* A gene was categorized as housekeeping if it was expressed in most tissues studied in the Gene Expression Atlas collection[49]

[48] http://NEIBank.nei.nih.gov/main/retina.shtml
[49] http://expression.gnf.org/cgi-bin/index.cgi

As expected, this list differs from that of the retSSH library and only three genes namely S-antigen (SAG), rhodopsin (RHO), and glutamate-ammonia ligase (GLUL) are abundantly found in both, the retSSH and the retNEIBank libraries.

3.2. Evaluation of the retNEIBank library

3.2.1. Estimation of the number of genes found in the library

Similar to the retSSH library, the 4913 clones were arbitrarily pooled into 40-clone plates and the number of new genes was plotted against the total number of clones. Since the number of new genes has not reached a plateau (Fig. 13), the analysis demonstrates that the number of sequenced clones is insufficient to estimate the number of genes present in the retina transcriptome. By extrapolation of the data plotted in Fig. 13B, it might be possible to identify five new genes per 40 clones sequenced even after sequencing more than 100,000 clones.

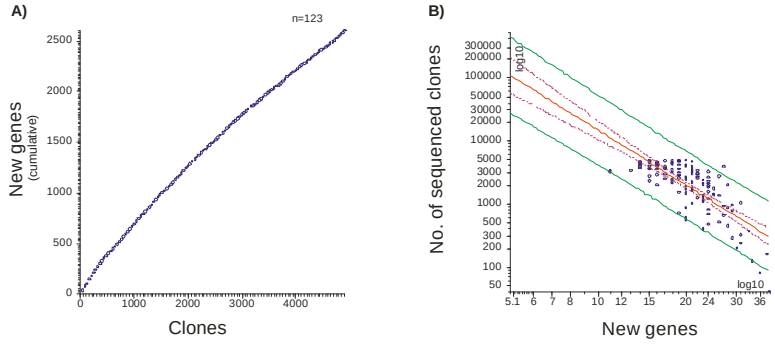


Fig. 13 Statistical analysis of the retNEIBank library
A) The cumulative number of new genes found after sequencing 4913 clones is depicted. As can be seen, the number of novel genes is still increases.
B) The number of novel unique genes found in each of the plates is shown as a function of the total number of sequenced clones. Using the data points the linear regression was calculated (red line). The pink line shows the 95% confidence interval of the regression line. The outer green line shows the 95% confidence interval of the individual points. Extrapolation of the number of novel genes found after sequencing 4913 clones randomly grouped in 'plates' of 40 clones each, predicts that even after sequencing 100,000 clones for each 40 sequenced clones it will be possible to identify five new genes.

4. Comparison of the retSSH and retNEIBank libraries

A comparison of the genes common to both libraries established that out of 2114 known genes from retNEIBank, 122 of them were also found in the group of 321 known genes from the retSSH library (Fig. 14). Interestingly even though many more clones were sequenced in the retNEIBank project, 199 genes from the retSSH are unique to this library. The fact that probably all of these transcripts are truly expressed in retina is supported by the fact that there is other retina library sequencing projects have also sequenced 246 of the 279 retSSH tissue-specific genes. For 16 genes no retina ESTs have been sequenced, but either fetal eye or retinoblastoma ESTs exist, thus for only 17 of the known tissue-specific genes from retSSH there is no previous evidence of retinal expression.

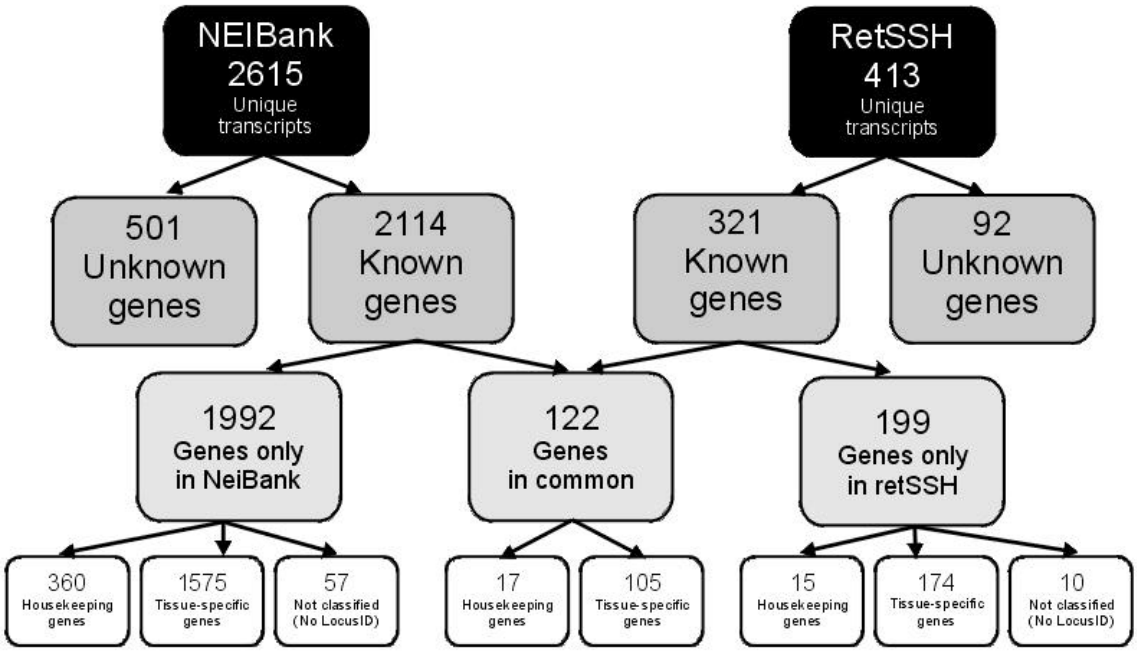


Fig. 14 Composition and comparison of the retNEIBank and retSSH libraries
While the 4913 clones of the retNEIBank library represent 2615 unique transcripts, the 1080 clones from the retSSH library are derived from 413 individual genes. Thus, on average there are 1.87 clones per retNEIBank transcript versus 2.60 clones per transcript in the retSSH library. The percentage of known genes in both libraries is very similar (81% and 77% for retNEIBank and retSSH, respectively). From these genes, 122 were found in both libraries. All genes found in the housekeeping list assembled from the HuGEIndex Gene Specific Expression database are counted as housekeeping genes. All those genes not found in this list or lacking a LocusID number were included in the other categories.

4.1. Analysis of the fraction of housekeeping genes found in the retSSH and retNEIBank libraries

A key issue to ensure that the retSSH library is actually enriched for retinal-transcripts is to demonstrate the absence of ubiquitous genes. Comparison of the retSSH and retNEIBank to the housekeeping gene list retrieved from the HuGEIndex Gene Specific Expression database[50] revealed the proportion of housekeeping genes in the retSSH library to be almost half of that for the retNEIBank (9.9% vs. 17.8%). If only the genes unique to each library are considered, the enrichment of tissue-restricted expression of the retSSH collection is actually even lower with only 15 out of 199 genes (7.5%) being ubiquitous.

5. Expression profiling of selected genes

5.1. Expression profiling by virtual Northern blot

For many years Northern blot analysis has been the gold standard to determine the expression of genes. But its use is hampered by the fact that large RNA amounts are needed and thus it is not suitable for serial expression analyses. On the other hand, reverse transcription of RNA with successive cDNA amplification has the disadvantage that it does not generate full-length transcripts.

[50] http://www.hugeindex.org/

However, nowadays it is possible to generate full-length cDNAs using a modified reverse transcription method. The cDNAs which are representative of the mRNA pool, can then be electrophoretically separated, transferred to a membrane, and hybridized with a labeled probe. This new technique utilizes the Southern blot technique, but since the membrane contains full-length cDNA, it is commonly referred to as virtual Northern blot. The advantage of this method is that it not only reduces the quantity of RNA required by a factor of at least 100, but also enables the determination of the full-length transcript size. However, in order to obtain cDNAs which are representative of the original sequence careful optimization of the PCR is necessary.

Because of these advantages, the methodology was chosen to confirm the expression and determine the transcript size of all genes identified as retina-specific or neuronal in the retSSH screening.

5.1.1. Optimization of the virtual Northern blot method

Since there is no standard protocol for the virtual Northern blot method, a number of optimizations were required before the methodology could be applied. The necessary optimizations included the determination of the best enzyme to be used for the secondary long-range amplification, the amount of first-strand product to be used, and number of cycles to be used for the second-strand synthesis.

In order to assess which polymerase was most suitable to use for the secondary amplification, an aliquot of 0.4 µl of first-strand cDNA was amplified in a final volume of 20 µl with either the Advantage *Taq* (BD Biosciences-Clontech) or the KOD HiFi *Taq* (Novagen) polymerases. After the 18th, 20th, 22nd, and 23rd cycle an aliquot of 5 µl was removed to asses the amplification efficacy after the given number of cycles. The aliquots were run on a 0.7% TBE/agarose/EtBr gel and transferred to a nylon membrane. As can be seen in Fig. 15, the amplification efficiency of the KOD polymerase was better, and therefore it was used in all future experiments.

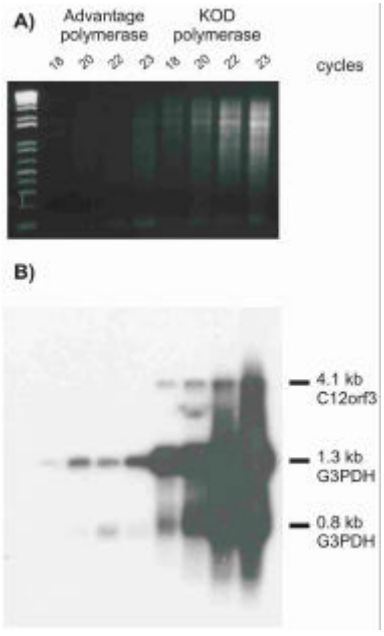


Fig. 15 Comparison of amplification of two polymerases
A) Electrophoretical separation of the products obtained by amplification of the same retina cDNA with the Advantage and KOD polymerases. The loaded 5 µl aliquots were taken from the reactions after the indicated number of cycles. The picture, taken after 1.5 h, shows that the Advantage polymerase was not as efficient as the KOD polymerase.
B) The products loaded on the gel pictured above were run an additional 1 hr and then blotted. The filter was hybridized simultaneously with two radio-labeled probes. The C12orf3 probe was used with the intention of assessing the efficiency of the polymerases to copy large transcripts. The G3PDH probe was hybridized in order to quantify the amount of product obtained at each stage with each enzyme. After stringent washing and an exposure of only 4 h, the signal from the KOD samples is still very strong, probably 5-fold stronger than that of the Advantage polymerase samples. The difference between the enzymes is not only quantitative, but also qualitative. The 4.1 kb C12orf3 product could only be detected in the samples amplified with KOD polymerase.

A major aspect to take into account when using an amplification procedure is the maintenance of the sample's original complexity. To reduce non-specific amplification and overrepresentation it is crucial to determine the amount of RNA to be used and to determine the optimal number of cycles for each sample. In our case four pilot amplification rounds with varying template quantities and number of cycles were needed until the optimal conditions for each tissue could be determined. This was done not only by observation of the electrophoretical separation of the reaction but also by hybridization of the blotted products with a radio-labeled probe for the housekeeping gene G3PDH (Fig. 16). Although brain cDNA was synthesized twice, the quality of the cDNA was bad both times and therefore this tissue had to be left out of the panel.

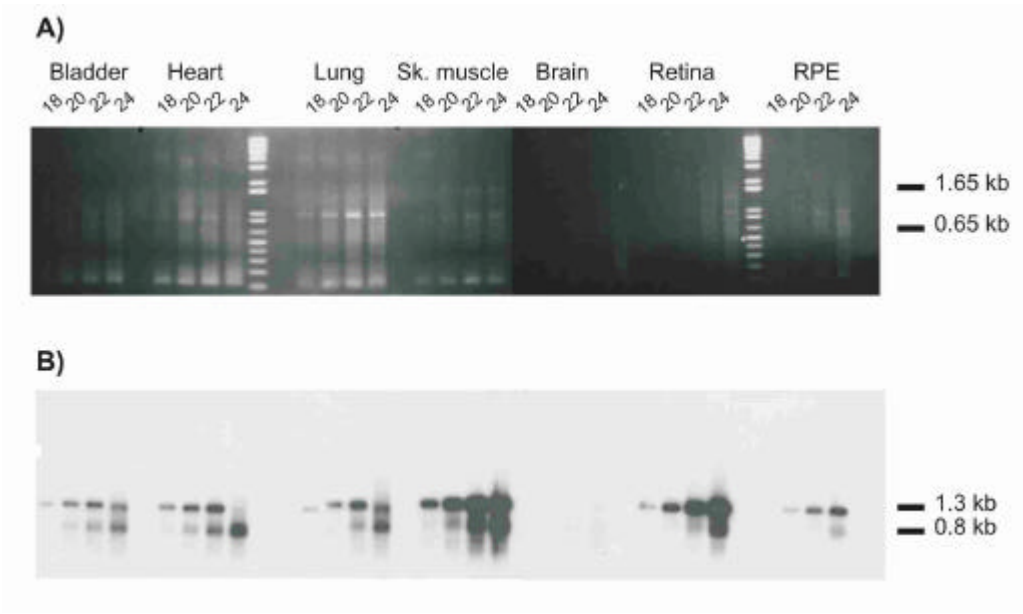


Fig. 16 Determination of the optimal template quantity and cycle number for each tissue
A) An aliquot of 5 µl taken after 18, 20, 22, and 24 cycles of the secondary PCR were electrophoretically separated on a 0.7% EtBr/TBE gel for 1.5h. As can be seen, each tissue has a particular band pattern determined by the genes most expressed in that tissue. After the picture was taken, the gel was run for another hour and then blotted to a charged membrane.
B) The membrane with the full-length transcripts of each tissue was hybridized with a G3PDH probe in order to estimate the quantity of transcripts present after a determined number of cycles. The probe labels two products, probably the gene and a pseudogene (Ercolani et al. 1988).

5.1.2. Assessment of the expression and transcript size of selected genes

A total of 36 genes from the retSSH project which were shown by RT-PCR to have a retina-specific or neuronal expression profile were selected for virtual Northern blot analysis (Table 19). The transcript size of 23 of the genes was reported in the literature. Nevertheless it was decided to use this method in order to confirm the size, to corroborate if the full-length cDNA is already cloned, and to investigate if there could be alternative splice variants in some of the tissues. A PCR product amplified with the primers listed in Table 19 was purified for each gene, labeled and hybridized to one of the 13 produced filters.

All the filters contained the same bladder, heart, lung, skeletal muscle, retina and RPE double-stranded cDNA, thus the uniformity between the different hybridizations is assured. It must be kept in mind, that since it is not possible to load exactly the same quantity of cDNA for each tissue, the expression profile has only a semi-quantitative value.

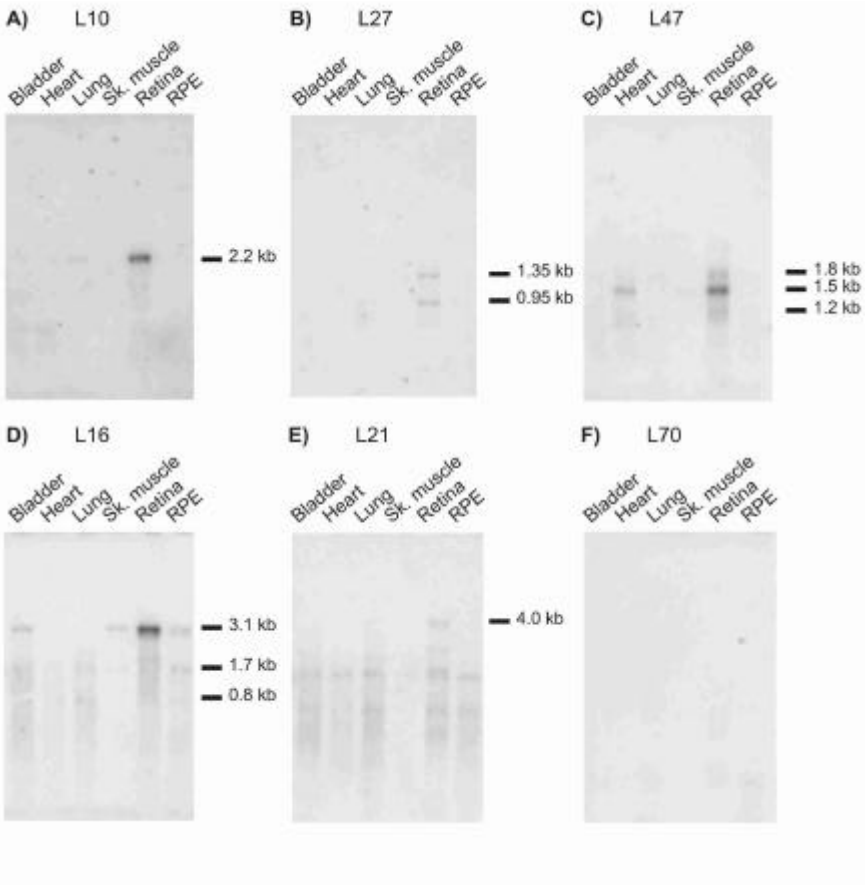


Fig. 17 Selected results of virtual Northern blot analyses
The name of the gene whose probe was hybridized is indicated above each picture. A) This gene is highly expressed in retina, although minor expression is also seen in bladder, heart, and lung. B) From what can be seen, there are two isoforms of L27 expressed only in retina. C) The heart- and retina-specific expression of L47 was corroborated but it seems that in retina there are additional splice variants. D) L16 is expressed at higher levels in retina than in bladder, skeletal muscle, or RPE. The 1.7 and 0.8 kb bands which may correspond to unspecific hybridization can be seen in all tissues. E) The L21 transcript reveals a large signal at 4.0 kb and the same 'unspecific' bands seen in 1D. F) No transcript could be visualized for approximately 30% of the genes, e.g. L70.

The expression and transcript size could be determined for two-thirds of the genes (Table 19). The transcript size of nine of these genes was previously unknown. At least one transcript of each of these 21 genes was expressed more abundantly in retina and/or RPE than in the other tissues (Fig. 17A, B, C, D, E).

Table 19 Summary of virtual Northern blot analyses

Lab ID	Reported transcript size	Size of bands (kb)	Signal intensity						Filter	Probe
			Bladder	Heart	Lung	Sk. muscle	Retina	RPE		
L02	At least 5614 bp	none	-	-	-	-	-	-	DS10	L2F/R
L03	4333 bp	none	-	-	-	-	-	-	DS19	L3F1/R
L05	At least 4614 bp	none	-	-	-	-	-	-	DS11	L5F/R
L10	2184 bp	2.2	+	+	+	-	++++	-	DS20	L10F/R
L14	2078 bp	none	-	-	-	-	-	-	DS14, DS22	L14F/R
L16	3116 bp	3.1	+	-	-	+	+++	+	DS10	L16F/R
L17	3220 bp	3.1	-	-	-	-	+	-	DS11	L17F/R
L18	2884 bp	4.0?	-	-	-	-	++	-	DS12	L18F/R
L20	2904 bp	2.7 2.0 1.4	- - ++	- - ++	- - ++	- - +	+ + ++	- - +	DS13	L20F/R
L21	6221 bp?	4.0	-	-	-	-	+	-	DS14	L21F/R
L23	At least 1749 bp	4.0 1.8	- -	- -	- -	- -	+ ++	- ++	DS15	L23F/R
L24	unknown	2.1 2.5	- -	- -	- -	- -	++ +	- -	DS21	L24F/R
L25	unknown	2.0	-	-	-	+	+++	-	DS10, DS11	L25F/R
L27	At least 801 bp	1.35 0.95	- -	- -	- -	- -	++ ++	- -	DS16	L27F/R
L28	4948 bp	none	-	-	-	-	-	-	DS17	L28F/R
L30	unknown	none	-	-	-	-	-	-	DS25	L30F1/R
L32	At least 1512 bp	?	-	-	-	-	-	-	DS19	L32F/R
L33	1635 bp	1.65	-	-	-	-	++	-	DS12	L33F/R
L35	unknown	none	-	-	-	-	-	-	DS10, DS23	L35F/R
L36	unknown	none	-	-	-	-	-	-	DS12	L36F/R
L37	unknown	0.4	-	-	-	-	+++	-	DS13	L37F/R
L38	unknown	1.1	-	-	-	-	++	-	DS18	L38F/R
L39	unknown	0.85 0.8	- -	- -	- -	- -	+++ +++	- -	DS14	L39F/R
L40	unknown	1.2 1.8?	- -	- -	- -	- -	+ +	- -	DS13	L40F/R2
L47	1425 bp	1.8 1.5 1.2	- - -	- ++ -	- - -	- + -	+ ++++ +	- - -	DS20	L47F/R
L48	unknown	1.05 0.85	- -	- -	- -	- -	+++ ++	- -	DS21	L48F/R
L50	unknown	3.2 2.2 0.65 0.6	- - ++ -	- - + -	- - - +	- - - -	+ + ++++ ++++	- - + +	DS15	L50F/R
L54	unknown	0.7	-	-	-	-	++++	+	DS22, DS26	L54F/R
L56	unknown	1.5	-	-	-	-	+++++	++	DS25	L56F/R
L63	1968 bp	none	-	-	-	-	-	-	DS20, DS22	L63F/R
L72	Approx. 3400 bp	none	-	-	-	-	-	-	DS23	L72F/R
L78	2636 bp	none	-	-	-	-	-	-	DS24	L78F/R
L86	At least 4782 bp	0.9	-	+	-	-	+	+	DS26	L86F/R
L88	At least 4352 bp	?	-	-	-	-	-	-	DS18, DS19, DS26	L88F/R
L92	unknown	none	-	-	-	-	-	-	DS27	L92F/R
L93	532 bp	0.6	-	-	-	-	+++	+++++	DS16	L93F/R

The number of + signals indicate the intensity of the observed bands. + = very weak signal, ++ = weak signal, +++ = strong signal, ++++ = very strong signal. The hyphen indicates absence of specific signal.

5.2. Expression profiling applying real-time quantitative PCR

The expression profiling of potential retina-specific or abundant genes described in the preceding sections was culminated by analysis using the real-time quantitative PCR (qRT-PCR) technology.

5.2.1. Optimization of qRT-PCR conditions

5.2.1.1. Optimization of SYBR Green concentration

SYBR Green is ideal for fluorescent detection because of its low cost and stability in the temperature extremes required for PCR reactions. However, it is a PCR-inhibitor, and therefore its concentration must be carefully determined. The optimal concentration is the one that has the highest possible increase in fluorescence combined with the lowest Ct-values. Generally SYBR-Green is not provided with an exact concentration; therefore the optimal concentration has to be determined empirically for each batch. This was accomplished by amplification with primers L40F2/L40R2 using serial dilutions of the SYBR Green 10,000x stock (Sigma-Aldrich Chemie GmbH, Munich, Germany) to 1:10, 1:100, 1:1000, 1:4000, 1:6000, 1:8000, 1:10,000, 1:12,000, 1:14,000, and 1:16,000. Analysis of the amplification curves (Fig. 18) determined that the ideal concentration is 0.5x (1:2000 dilution).

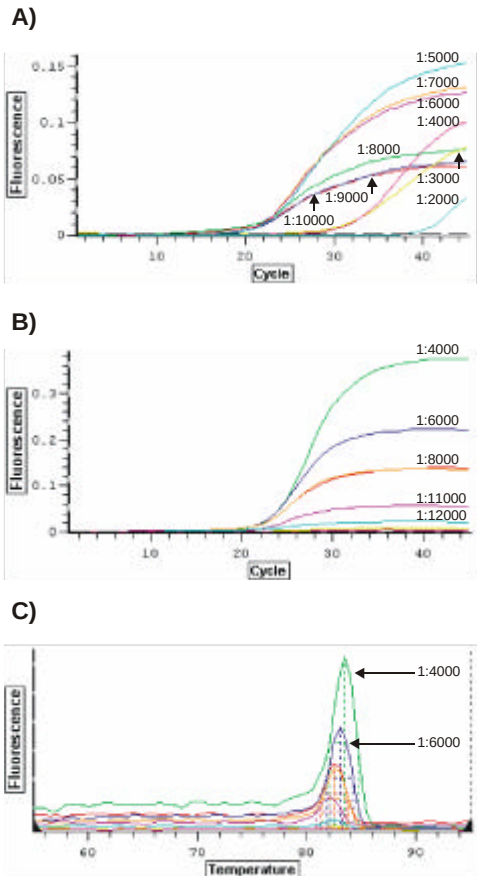


Fig. 18 Optimization of SYBR Green concentration
A) Primers L40F2/L40R2 were added to a master mix containing all components except SYBR Green. The mix was distributed in nine wells and varying dilutions of SYBR Green (from 1:2000 to 1:10,000) were added to each. The inhibitory effect of SYBR Green is clearly seen, as almost no product is amplified if the 1:2000 dilution is used. On the other hand, if the SYBR Green concentration is too low, the signal intensity is low. Therefore, for this batch, a dilution of 1:5000 was considered optimal.
B) A second batch of SYBR Green was optimized, and in this case, the optimal efficiency is attained with the 1:4000 dilution.
C) Melt curve of the second SYBR Green optimization.
These reactions were carried out in an Opticon2 detection system.

5.2.1.2. Primer design

The use of SYBR Green in qRT-PCR requires extremely careful design of primers since the dye binds to all double-stranded DNA products. Aside from the general rules for primer design such as GC-content and oligonucleotide length it is necessary to comply with other requirements. First the amplified product size should be between 75 and 150 bp long in order to favour consistency in the amplification efficiency. Second, the primers should span an exon-exon junction, but not more than 5 bp of the 3'-end of the intron-spanning primer should be located in the other exon. This is important to prevent amplification of genomic DNA possibly contaminating the RNA samples which would lead to an overestimation of the quantity of template. Third, once both primers have been designed, it is vital to analyze if they can interact with each other generating cross dimers which may be extended. Finally, the primer Tm temperature should be as high as possible in order to allow annealing at the highest possible temperature (ideally around 64°C) so that primer-dimer formation will not be possible.

A total of 140 primers (Table 37, Appendix) were designed for the analysis. In a few cases more than one primer pair had to be selected, as the original primers could not be optimized.

5.2.1.3. qRT-PCR optimization

For successful qRT-PCR a set of rules has to be followed to obtain the highest possible efficiency while maintaining the primer-dimer formation at a low concentration. This meant assessing as many as 10 different conditions for some oligonucleotide combinations. The main factors varied in the optimization process were the $MgCl_2$ concentration and the annealing temperature. Additionally, in cases were primer-dimer formation was unavoidable, an extra step in the cycling was incorporated in order to measure the fluorescence at a temperature where the primer-dimer signal was minimal but the specific-signal fluorescence was still present.

The addition of SYBR Green usually demands a higher $MgCl_2$ concentration in order to stabilize the PCR product. Therefore, the final $MgCl_2$ concentration is generally between 1.5 and 4 mM. An excess of $MgCl_2$ may lead to non-specific priming, whereas too little will result in sub-optimal polymerase activity. For all PCRs done in this study, optimal results were obtained when the final $MgCl_2$ concentration was between 2 and 3 mM (Fig. 19).

An advantage of the iCycler system is the multiple fluorescence measure for each acquisition step which allows the determination of the optimal duration of the elongation step. Since the products are small, we observed that in all cases in which detection was done at a temperature higher than 72°C to avoid measuring primer-dimer signal, the elongation step at 72°C was unnecessary (Fig. 19C). The annealing temperatures used in the study ranged from 57°C to 67.5°C (Fig. 20).

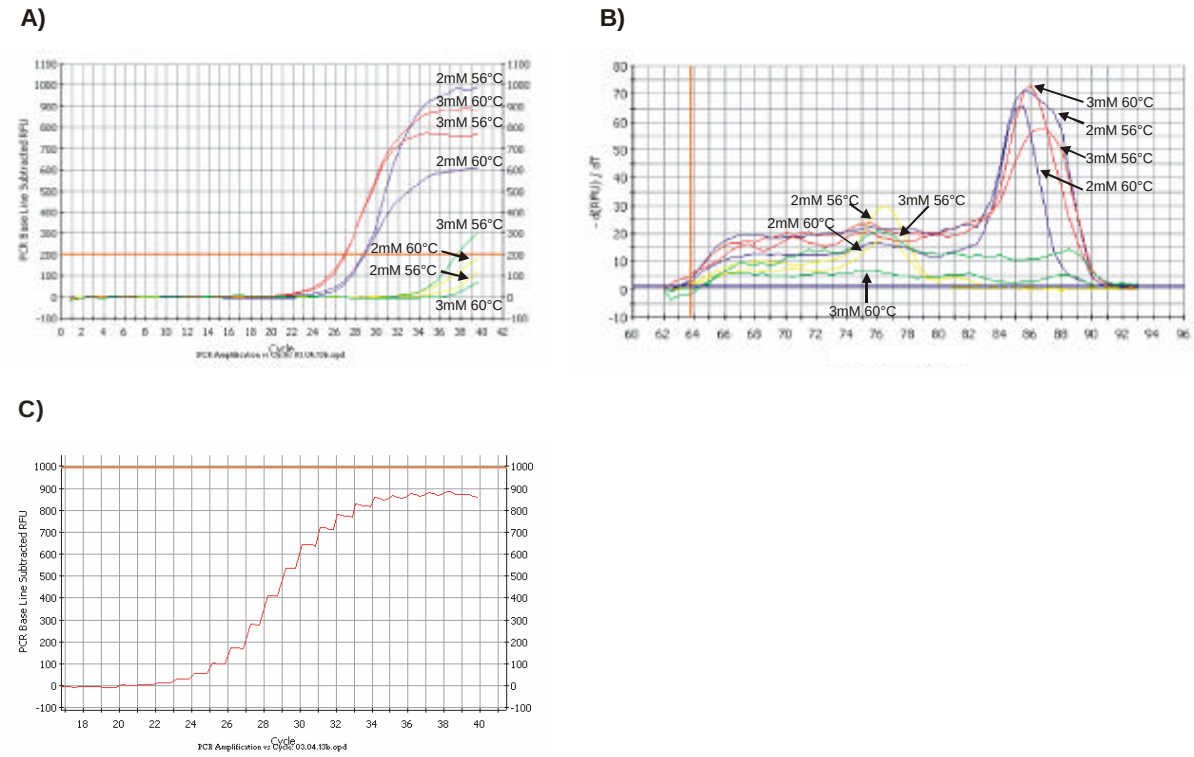


Fig. 19 Effect of $MgCl_2$ concentration and temperatures

A) Amplification of retina cDNA using primers L25F4 / L25R4 in either a 2 or 3 mM $MgCl_2$ buffer. The annealing temperature (Ta) was set to either 56°C or 60°C for each combination. Negative control samples (yellow and green lines) were also simultaneously run using the same conditions. As can be observed the annealing temperature does not alter the Ct as much as the $MgCl_2$ concentration. In the more stringent 2 mM reactions the threshold is reached two cycles later. Although the effect of the annealing temperature and $MgCl_2$ concentration may not be so important in samples with template, in those without template the magnitude of unspecific signal is greatly influenced by the conditions. As expected, the less stringent condition (3 mM $MgCl_2$, Ta = 56°C) produces the greatest amount of unspecific signaling.

B) The melting curve analysis of the products amplified in A) was used to select the optimal condition for this primer pair. As can be observed, the 2 and 3 mM $MgCl_2$ 56°C reactions produce broad curves which are probably due to the presence of more than one product. The negative control curves reveal the presence of primer-dimers in all conditions except combination of 3 mM $MgCl_2$ and 60°C. Therefore, even though these are not the most stringent conditions, they were selected for all future amplifications of the gene.

C) The actual values measured for the amplification curve using 3 mM $MgCl_2$ and 60°C is shown. Each 'step' shows the variation in the fluorescence during the 8 sec in which it was measured. As can be seen, the elongation step can be skipped, since no additional copies of the transcript are synthesized at this step.

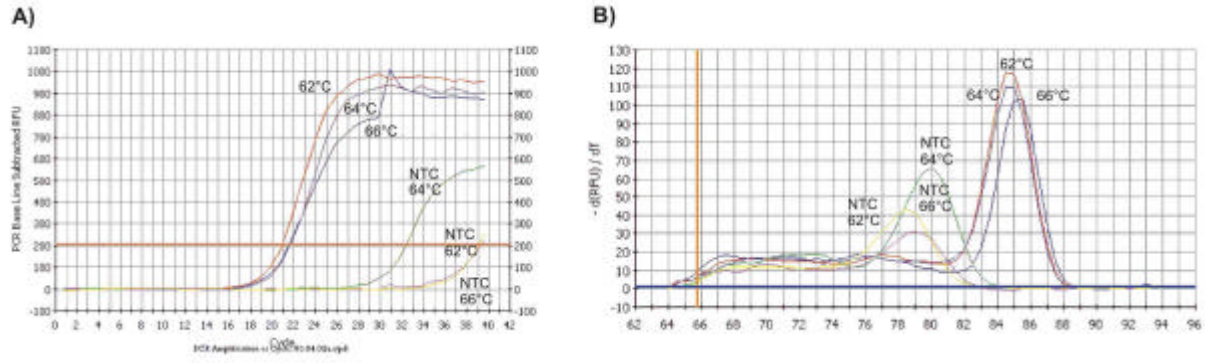


Fig. 20 Effect of different annealing temperatures

A) The amplification of retina cDNA with primers L47F/L47rR1 in a mix with a final $MgCl_2$ of 2mM and annealing temperatures of 62°, 64°, and 66°C. The annealing temperature does not have much influence on the amplification efficiency, but has a marked effect in the production of primer-dimers. As can be seen, the production of primer-dimers does not always follow strict rules. This is demonstrated by the fact that more primer-dimers are generated if an annealing temperature of 64° instead of 62°C is used. The sudden peak in the 66°C curve is artefactual and due to an air bubble.

B) Melting curve analysis of the reaction depicted above. The X-axis shows the temperature at which the products melt. The peak of the curve indicates the temperature at which half of the product is found as single-strand and therefore emits no fluorescence. For the expression analysis of this gene the reaction was done at an annealing temperature of 66°C and the reading was done at 82°C so that no dimer signal would be measured.

5.2.1.4. Determination of the qRT-PCR reaction efficiency

To accurately determine the relative quantity of a template, it is necessary to investigate the efficiency of amplification for the primer pair used. Several approaches have been proposed to accomplish this. A selection of the most commonly used includes the classical calibration dilution curve and slope calculation, the increase in absolute fluorescence method (linear regression) (Pfaffl 2001), absolute fluorescence window-of-linearity (Ramakers et al. 2003), and the experimental sigmoidal fit (Tichopad et al. 2003). Two of these methods, the absolute fluorescence window-of-linearity and the calibration dilution curve, were applied and it was finally decided to follow the latter.

The calibration curve was generated by five serial fourfold or fivefold dilutions of retina cDNA. The threshold cycles obtained for triplicates of each dilution were plotted against the log of the sample concentration (Fig. 21). For the sample concentration arbitrary numbers representing the four-fold difference were used, since the concentration is unknown and not relevant. The slope of the line linking all points was then used to calculate the efficiency (E), with $E = 10^{[-1/slope]}$. An efficiency value of 2 means the amount of product is doubled after each cycle. If the efficiency has a smaller value there will not be a duplication of product quantity at each cycle. The efficiencies of all primer pairs used is detailed in Table 37 (Appendix). An E value of 2 corresponds to 100% efficiency whereas E= 1.5 translates to 75%. The efficiency of the primer combinations of this study ranged from 79% to 104%, with an average of 93%.

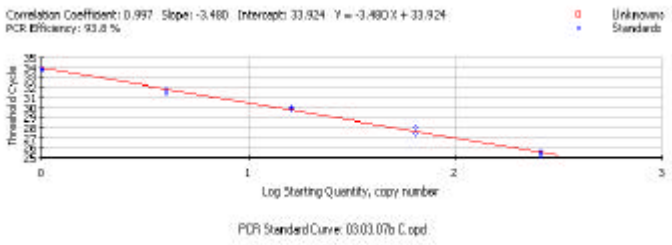


Fig. 21 Standard dilution curve used to calculate PCR efficiency
Retina cDNA was diluted in five consecutive 1:4 dilutions in order to construct the dilution curve. Each dilution was amplified in triplicate with primers L54rF2/L54rR2 which amplify a 149 bp product. To construct the curve and establish the slope value, the threshold cycle of each sample was plotted against the logarithmic starting concentration. The slope value of -3.480 translates to a PCR efficiency of 94%.

5.2.1.5. Normalization procedures

In this study, the quantification of the gene expression was accomplished by normalization to standard genes. The ideal standard should be expressed at a constant level among the different tissues under investigation, at all stages of development, and should be unaffected by any experimental treatment. In addition, the control genes should be expressed at roughly the same level as the gene under study. Obviously, there is not a single transcript complying with these requirements and therefore it is necessary to do a careful selection to choose the optimal normalization genes. Historically, the most commonly used genes have been glyceraldehydes-3-phosphate-dehydrogenase (G3PDH), ß-actin, and ribosomal RNA (Suzuki et al. 2000). For several reasons, none of these genes are actually suitable for the normalization process and we therefore selected other more appropriate genes.

To validate the stability of expression for a given gene it is necessary to measure the stability. Therefore, to attain an acceptable standard in the normalization procedure of our study, the expression of several putative housekeeping genes was investigated in all the tissues included in the panel. In order to cover all ranges of expression we included a. highly-expressed genes: ß-actin (ACTB) and beta-2-microglobulin (B2M); b. moderately-expressed genes: ribosomal protein L13a (RPL13A) and succinate dehydrogenase complex, subunit A (SDHA); c. low-expression genes: hypoxantine phosphoribosyl-transferase I (HPRTI), hydroxymethyl-bilane synthase (HMBS), beta-glucuronidase (GUS) and TATA box binding protein (TBP). The PCR efficiency of each primer combination was calculated as already described (Table 37) and the expression was tested in triplicate for each cDNA. In order to compare the results of all genes, the Ct values of each tissue and gene were converted to a quantity using the formula $E^{(\text{minimal Ct} - \text{Ct of investigated tissue})}$, where E is the efficiency of the PCR and minimal Ct is the lowest threshold value of the analyzed samples.

To select the most stable housekeeping gene from this set of tested genes, we used the geNorm VBA applet[51] for Microsoft Excel (Microsoft, Unterschleissheim, Germany) developed by Vandesompele et al. (2002). This approach assumes that minimally regulated, stably expressed genes stay in a constant ratio to each other. Therefore an internal control gene-stability measure M, which is the average pairwise variation of a particular gene with all other control genes, is calculated for each tissue. Thus, the genes with the lowest M values have the most stable expression. It also calculates a gene expression normalization factor for each tissue based on the geometric mean of the housekeeping genes. An example of the values obtained for each of the tissues of panel T1 and the calculated normalization factor is shown in Table 20. The selection of the most appropriate genes and calculation of the normalization factor was done for each of the three panels used. Since each panel was reverse transcribed at different times, logically the normalization coefficient for each tissue varied, but the stability order of the genes was constant. Thus, for our panel the genes ranked in order of their expression stability are RPL13A (most stable), GUS, ACTB, TBP, SDHA, and B2M (least stable).

Table 20 geNorm calculation of the most stable housekeeping gene and normalization factors (NF)

Tissue	GUS	SDHA	B2M	TBP	ACTB	RPL13A	NF
Basal ganglia	0.128969463	0.125225327	0.1163051	0.3721601	0.1770308	0.2006912	**0.5975**
Cerebellum	0.40433522	0.295639021	0.1042138	0.8131164	0.2109579	0.4381038	**1.0942**
Occipital cortex	0.235862651	0.340011819	0.0913518	0.8914251	0.4493446	0.3241911	**1.0971**
Retina	0.470202186	0.398935154	0.0311542	1	0.1093061	0.2742508	**0.8275**
RPE	0.099569733	0.033502232	0.0701941	0.079775	0.0736739	0.2098469	**0.2844**
Bladder	0.635873588	0.251972684	0.5407895	0.8513709	1	1	**2.2655**
Distal colon	0.230831929	0.071575295	0.4438292	0.2460567	0.312987	0.5008408	**0.8963**
Heart	0.312163429	1	0.3089536	0.4903725	0.1421887	0.5854765	**1.3901**
Lung	1	0.190497354	1	0.6461338	0.7520751	0.7824234	**2.2586**
Stomach	0.40433522	0.146926613	0.1652548	0.2890124	0.1890618	0.6545532	**0.9292**
M	**1.0092**	**1.4162**	**1.4670**	**1.1712**	**1.1402**	**0.9947**	**1.1997**

Another issue which has to be considered is the number of genes which should be included for the normalization. It is recommended to use the three most stable control genes and include more genes

[51] http://allserv.rug.ac.be/~jvdesomp/genorm/index.html

until the addition of a new gene has no significant contribution to the newly calculated normalization factor. The best way to accomplish this is by calculating the pairwise variation between two sequential normalization factors for all samples within the same tissue panel. In our panels, the inclusion of an additional gene always lowered the variation (Fig. 22), thus, all tested genes were included for the calculation of the normalization factor.

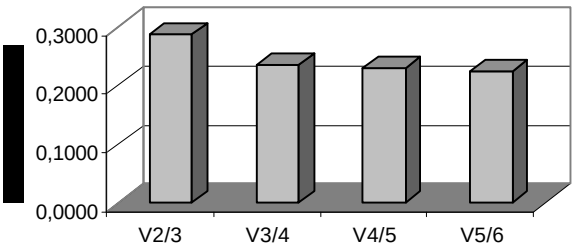


Fig. 22 Determination of the optimal number of control genes for normalization
Pairwise variation analysis between the normalization factor obtained by using two and three genes (V2/3), three and four (V3/4), and so forth.

5.2.2. Quantitative expression profiling of selected genes by qRT-PCR

After optimization of the conditions discussed above, the expression of 52 genes known to be expressed specifically or abundantly in the retina was determined by qRT-PCR. A list of the genes amplified and their lab ID is detailed in Table 21.

Table 21 Genes whose expression profile was determined by qRT-PCR

Lab ID	Primer design based on sequence
A004	CAMTA1
A017	ERO1LB
A059	PSMD11
A084	KIAA1796
A085	BC042097
A106	FLJ13993
A109	KIAA1263
A111	BC016878
A126	C12orf3
A150	AK054981
A165	CHD1
A166	C1orf32
A168	ORC2L
A169	AA057097, in intron of ADAMTS18
A177	DKFZp761D221
A203	STK35 alternative splice?, AL844428
A205	BC035234
A206	AK056484
A211	SF3B3
A213	CRYPTIC
B001	FLJ31564
B015	C14orf
B030	AK091467
L02	CLASP2
L05	KIAA1380
L10	C14orf129

Lab ID	Primer design based on sequence
L11	KIAA1576
L14	FLJ33282
L16	PLCD4
L17	SSX2IP
L18	FLJ13305
L21	EKI1
L23	C20orf103
L24	SLC1A2
L28	SV2B
L33	BC029061
L35	-
L36	DKFZp434C0631
L37	BM668448
L38	-
L39	C4orf11
L40	ZPBP
L47	MGC14816
L48	-
L52	BC040189
L54	-
L56	-
L63	DKFZp547C176
L72	H2AV variant 2
L78	FLJ30499
L88	KIAA1579
L93	DAPL1

- : Sequence is not publicly available.

The expression of each gene was determined in one of three panels identified as panel O, H1, and T1. The panels contained cDNAs from tissues of ectodermal (occipital cortex, cerebellum, basal ganglia, retina, and RPE), mesodermal (heart, bladder), and endodermal (lung, bladder, stomach)

embryological origin. The first panel also contained skeletal muscle cDNA, but since most primers amplified artefactual products from this cDNA, it was eliminated from future panels and replaced by stomach cDNA. As already explained, primers were selected for each gene, optimized, and the amplification efficiency was calculated by using the dilution curve method. The results obtained in each tissue were normalized to the values calculated for each of the three panels (Table 22).

Table 22 Normalization factors for all tissues and panels

Tissue	Panel O	Panel H1	Panel T1
Basal ganglia	0.3186	0.1750	0.1547
Cerebellum	0.5856	0.3098	0.2833
Occ cortex	0.5331	0.2659	0.2846
Retina	0.2351	0.1730	0.2137
RPE	0.1281	0.0774	0.0731
Bladder	0.3705	0.5081	0.5856
Distal colon	0.3427	0.2758	0.2308
Heart	0.4810	0.3997	0.3581
Lung	0.5221	0.5099	0.5843
Stomach	n.i.	0.2357	0.2407

n.i. Stomach was not included in this panel

The expression analysis was done using duplicate or triplicate samples for each cDNA template. The PCR product identity was established by melting curve analyses and electrophoretical analysis of at least one sample of each reaction. The results were analyzed after baseline-subtraction but without curve fitting. For most cases the optimal threshold determined by the software was accepted, only in cases were it seemed necessary to do otherwise, it was manually set. If two of the duplicates had the same Ct value while the third was in discordance, then the third data point was removed for the quantitative calculations.

The exported Ct values for each sample were processed in a specially designed Excel (Microsoft, Unterschleissheim, Germany) table developed by ourselves. The use of macros and visual basic for applications (VBA) greatly automated the generation of results and decreased the possibility of calculation errors. Although there are worksheets designed for this task (e.g. BestKeeper©), they are programmed to normalize using only a single housekeeping gene and were therefore not suitable for our study.

	Tissue	Ct	Conversion	Average	SD
TBP	Cerebellum	26,1	0,501775133	0,4705173	0,0541
1,993	Cerebellum	26,1	0,501775133		
	Cerebellum	26,4	0,408001608		
	Occ cortex	25,1	1	0,8404924	0,1477
	Occ cortex	25,4	0,813116437		
	Occ cortex	25,6	0,708360877		
	Retina	25,3	0,871167824	0,7626299	0,094
	Retina	25,6	0,708360877		
	Retina	25,6	0,708360877		
SDHA	Basal ganglia	28,4	0,085673691	0,0874375	0,0031
1,821	Basal ganglia	28,3	0,090965264		
	Basal ganglia	28,4	0,085673691		
	Cerebellum	26,9	0,210508255	0,1914692	0,0232
	Cerebellum	27	0,198262702		
	Cerebellum	27,3	0,165636746		
	Occ cortex	26,7	0,237315088	0,2283795	0,0155
	Occ cortex	26,7	0,237315088		
	Occ cortex	26,9	0,210508255		
	Retina	26,9	0,210508255	0,2170092	0,0092
	Retina	26,8	0,223510145		
	Retina	26,8	0,223510145		

Calculation of normalization factor (NF) and standard deviation (SD) of NF

Tissue	TBP	SDHA	NF	SD NF
Cerebellum	0,84049244	0,191469234	0,309817	0,016161653
Occ cortex	0,76262986	0,228379477	0,2659238	0,007293112
Retina	0,96668186	0,2170092	0,17295575	0,003520064

Fig. 23 Normalization factor calculation worksheet

The only values that have to be entered are the Ct values for each tissue, the PCR efficiency values for the housekeeping genes (e.g. TBP, SDHA), and the tissues included in the assay. The cells were the user has to input information have a light grey background. All other information is calculated or copied automatically. The cells that are filled automatically have a dark-grey shade.

In a first worksheet the Ct values for each qRT-PCR reaction of the housekeeping genes are entered (Fig. 23). The macros then automatically calculate the normalization factor and standard deviation for each tissue using the formulas published for geNorm[52]. A worksheet is then created for each gene. After entering the gene name, name of file of the expression analysis, slope of the efficiency curve, and Ct values the macros automatically name the worksheet, calculate the PCR-efficiency, convert the Cts in quantities (using a range of 0 to 1), compute the standard deviation, calculate the relative expression in each tissue by applying the normalization factor, and rescale the relative expression values so that the tissue with the least expression has a value of one. A graphical representation of the results is also generated.

Gene	L16
File	030401f
Slope	-3,712
Efficiency	1,85950158

	Ct	Average Ct	Conversion	Average	SD
Cerebellum	30,3	30,56666667	0,012226003	0,0104472	0,001650332
Cerebellum	30,6		0,010149988		
Cerebellum	30,8		0,008965741		
Occ cortex	29,4	29,4	0,021366893	0,0214767	0,002659318
Occ cortex	29,2		0,024189155		
Occ cortex	29,6		0,018873918		
Retina	23,2	23,36666667	1	0,9067059	0,113556636
Retina	23,3		0,939853893		
Retina	23,6		0,780263658		

EXPRESSION

	Expression	SD
Cerebellum	0,033720693	0,006
Occ cortex	0,080762442	0,01
Retina	5,2424153	0,665

EXPRESSION RESCALED

	Expression	SD
Cerebellum	1	0,166
Occ cortex	2,395041017	0,304
Retina	155,4658246	19,73

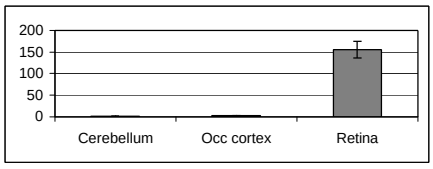


Fig. 24 Gene expression worksheet
After input of the gene ID, name of the file containing the raw data, value of the efficiency curve slope, and Ct values (in the light-grey cells), the macros calculate the average Ct values, transform the Ct values to a numeric scale, averages the values and calculates the standard deviation (SD). The average value is then divided by the normalization factor (NF) calculated for each tissue in the normalization factor worksheet. Likewise, the SD is also calculated taking into account the SD of the NF. The result is then rescaled so that the tissue with the lowest expression has a value of 1. The content of the black and the dark-grey cells is automatically filled in or calculated. The graphical output is also generated automatically.

In the following figure, the relative expression of each tested gene is presented. The identification found on the top of the graph corresponds to the lab intern ID of each gene. A correlation of this ID with the actual gene name is found in Table 21. Please note that the relative expression values of the Y-axis vary between the genes. There were also results that contained spurious signals and therefore the measured expression is overestimated or could not be calculated. These datasets are depicted with black bars in order to point out a possible overestimation of the expression. The genes for which only a partial quantification is available are A085, A109, A177, L23, L37, L38, L39, L48, L54, L63,

[52] http://allserv.rug.ac.be/~jvdesomp/genorm/index.html

A166, and A169. There were also genes for which there was absolutely no expression of the gene in a particular tissue, and thus there is no reported Ct value (L35, L40, L56, and A106). In such cases a fictitious value of 50 needs to be entered in the worksheet.

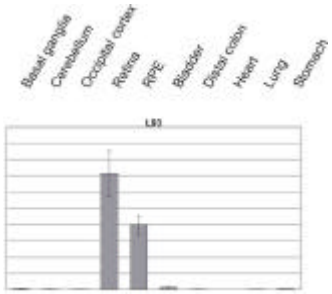


Fig. 25 Quantitative expression profile of 52 genes (next three pages)

The quantity of transcript present in each of the tissue detailed above the graphs was determined by qRT-PCR. The expression value obtained for each tissue was normalized by comparison to six control genes. For the graphical representation the normalized expression values were converted to represent the relative expression in relation to the tissue with the lowest expression (calibrator) and displayed as the fold change. The black bars indicate that the expression in that tissue is probably overestimated because the fluorescence of more than one product was measured. The Lab ID of each gene is indicated above the graph.

Figure 25 Quantitative expression profile of 52 genes (contd)

Figure 25 Quantitative expression profile of 52 genes (contd)

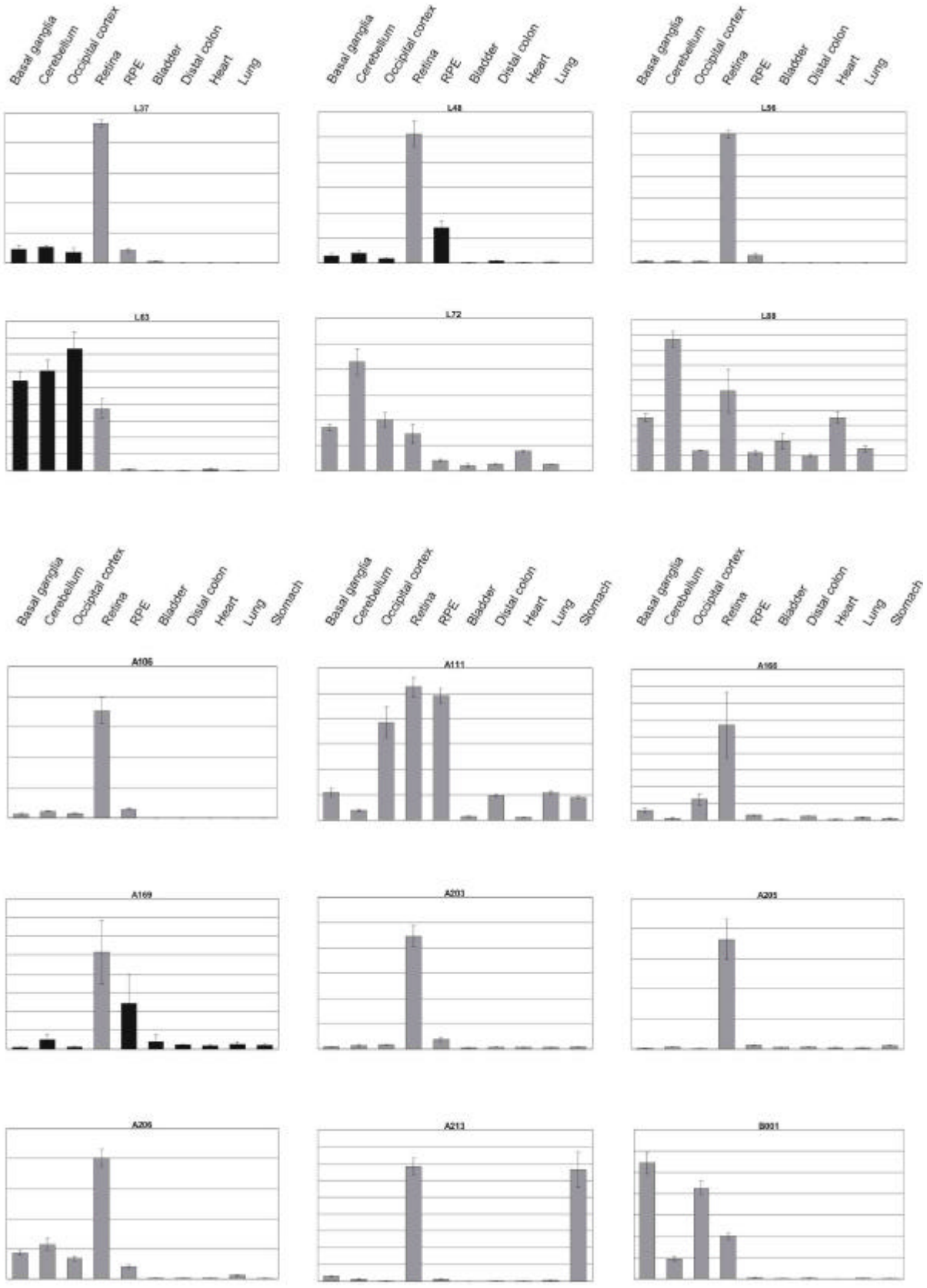
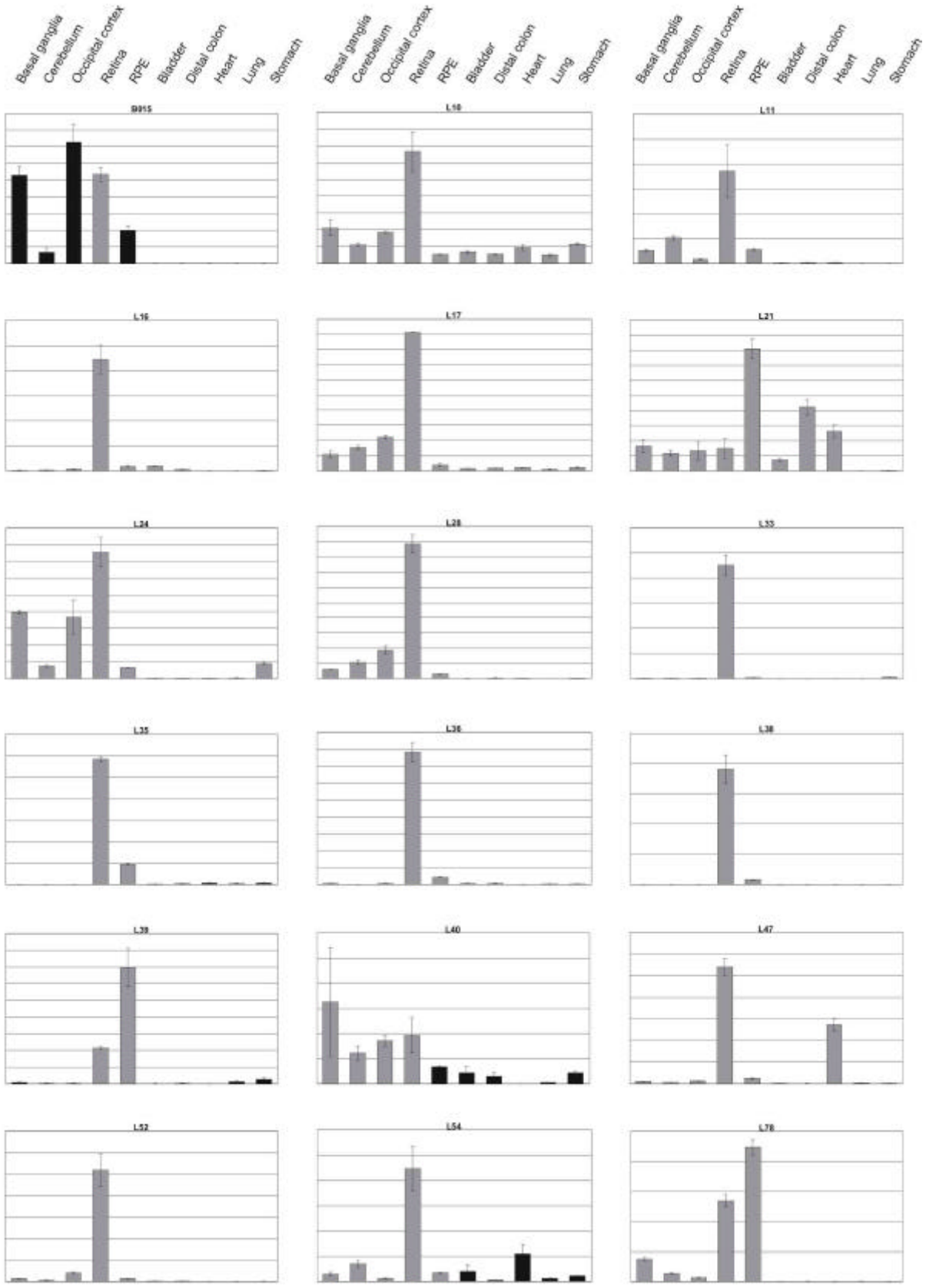


Figure 25 Quantitative expression profile of 52 genes (contd)


Basal ganglia
Cerebellum
Occipital cortex
Retina
RPE
Bladder
Distal colon
Heart
Lung
Stomach
B915
L10
L11
L16
L17
L21
L24
L28
L33
L35
L36
L38
L39
L40
L47
L52
L54
L78

6. Cloning and characterization of genes preferentially expressed in the retina

6.1. Cloning and characterization of C7orf9 (A129)

6.1.1. Assembly of the cDNA sequence of C7orf9

Partial sequences corresponding to C7orf9 were found in Hs. cluster 60473 which contained the 3'-end sequence of two Soares retina N2b4HR clones (ze37g05 and ze34f06). A search in the dbEST database revealed that approximately 450 bp of the 5'-end of these cDNA clones were also available. To isolate novel clones a retina cDNA library, constructed in the TriplEx2 vector, was screened. The screening, done with a radio-labeled 199 bp DNA fragment obtained by PCR amplification with primers A129F and A129R (Fig. 26 and Table 36), identified fourteen positive clones (129-3-1, 129-4, 129-5-2, 129-7, 129-8, 129-8-1, 129-12-4, 129-13-2, 129-13-3, 129-14, 129-15-1, 129-17, 129-18-2, and 129-27A). The inserts of the clones, which ranged from 0.5 to 1.2 kb, were isolated and partially or completely sequenced. Sequence assembly yielded an 1190 bp transcript, termed C7orf9 (GenBank acc. no. AF440392), with an open reading frame (ORF) of 591 bp and a potential start codon (ATG) located 48 bp downstream of the 5'-end of the cDNA (Fig. 26). To clone the complete 5'-end of the gene, a first strand cDNA synthesis was performed using the gene-specific oligonucleotide primer A129R (Fig. 26 and Table 36). This was followed by 5'-RACE amplification with the gene-specific reverse primer A129R4 and AUAP (Fig. 26 and Table 36). To increase the amount of the specific product, a 1:100 dilution of the original PCR was re-amplified using primers A129R5 and AUAP (Fig. 26 and Table 36). The amplified products were sequenced, but they did not contain further upstream sequences suggesting a comprehensive coverage of the most 5'-region of the transcript. Three putative polyadenylation signals were identified at positions 801 bp, 908 bp, and 1109 bp (Fig. 26). BLASTN searches[53] with the full-length cDNA sequence revealed significant identity to two independently isolated human cDNA transcripts (GenBank acc. nos. AB040290 and AF330057) which were made public at a later time point. Sequence AF330057 contains only part of the coding sequence; AB040290 contains the complete coding sequence but does not contain either the 5'- or 3'-UTR.

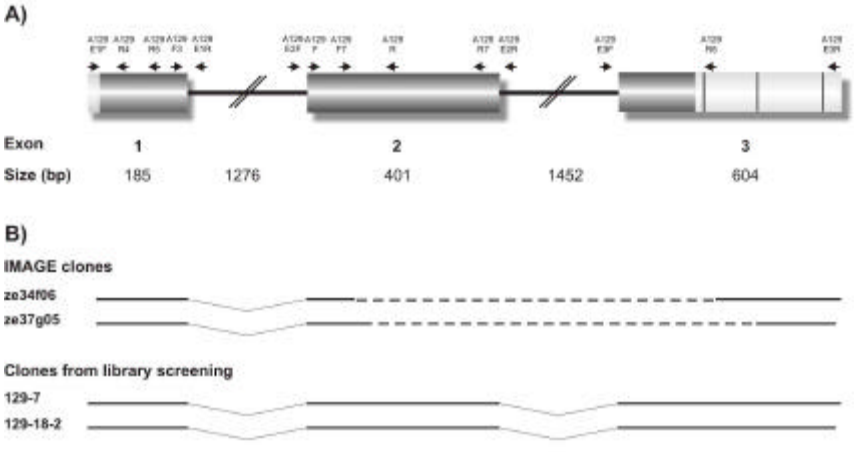


Fig. 26 Genomic structure and representative clones of C7orf9
A) The three coding exons of C7orf9 are represented by shaded boxes. Light grey boxes represent the 5'- and 3'-untranlated regions (UTRs). Three potential poly-adenylation sites are depicted as vertical black bars in the 3'-UTR. The name and relative positions of oligonucleotide primers utilized in the study are shown.
B) The consensus cDNA sequence was assembled from I.M.A.G.E. retina clones ze34f06 and ze37g05 and sequencing of cDNA clones retrieved from retina cDNA library screenings (clones 129-7 and 129-18-2). The dotted sequence represents unknown sequence.

[53] http://www.ncbi.nlm.nih.gov/BLAST/

6.1.2. Genomic structure of C7orf9

To determine the genomic structure of C7orf9, the cDNA sequence was aligned to the working draft sequence of BAC clone CTB-136N17 (GenBank acc. no. AC004129). The three exons spanning approximately 4 kb of genomic DNA from chromosome 7p15.3 are separated by two intervening sequences of 1275 bp and 1452 bp (Fig. 26). All identified donor and acceptor splice site sequences conform to the GT-AG rule (Penotti 1991 and Berg and von Hippel 1998).

Table 23 Detailed information about the exonic/intronic structure of C7orf9

Exon		3'-Acceptor Splice Site[a]		5'-Donor Splice Site[a]		Intron	
No.	**Size (bp)**	**Sequence**	**Score[b]**	**Sequence**	**Score[b]**	**No.**	**Size (bp)**
1	185			GAGgtaagt	0.87	1	1276
2	401	tcatttctaattatagCCTA	7.53	AAGgtaaat	2.68	2	1452
3	604	actgcttacattttagGAGA	7.56				

[a] Exonic and intronic sequences in upper and lower case letters, respectively.
[b] Score of donor/acceptor splice site. According to published data (Berg and von Hippel 1998 and Penotti 1991) 99% of sites have a score of 0-11 (donor) or 0-20 (acceptor). Scores were calculated using the spreadsheet created by Christian Sauer, 2001.

6.1.3. Expression analysis of C7orf9

To examine the expression of C7orf9, RT-PCR analysis was performed with primers A129F3 and A129R (Fig. 26 and Table 36). A 244 bp transcript was amplified exclusively from human retina but not from the other 19 tissues tested (Fig. 27 A). Northern blot hybridization with the above-mentioned fragment labeled with a^{32}P-dCTP confirmed this pattern of expression (Fig. 27 B). Two bands of approximately 0.8 kb and 1.2 kb were labeled only in retina, suggesting that two of the three hypothetical polyadenylation sites, namely those at 801 bp and 1109 bp, may be used. In agreement with this finding, expression of the rat orthologue was seen to be restricted to hypothalamus and eye (Hinuma et al. 2000).

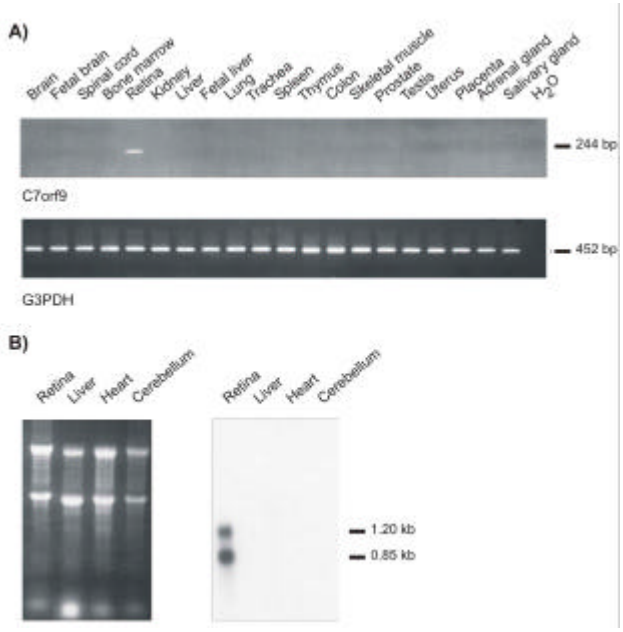


Fig. 27 Expression analysis of C7orf9
A) RT-PCR amplification of a 244 bp fragment using primer pair A129F3/A129R was only observed in the retina cDNA. No amplification was observed in the remaining 19 tissues. As a positive control all cDNAs were amplified with primer pair G3PDH-ex8F/G3PDH-ex9R which anneal to the ubiquitous gene glyceraldehydes-3-phosphate dehydrogenase (lower picture).
B) RNA from four tissues was separated on a formaldehyde agarose gel and stained with ethidium bromide (left). The resulting Northern blot was probed with an a^{32}P-dCTP-labeled PCR product amplified by A129F3 and A129R. Two signals at approximately 1.2 kb and 0.85 kb were detected in retina, but not in the other tissues tested.

6.1.4. *In-silico* analyses of the putative C7orf9 protein

Translation of the C7orf9 cDNA results in a putative 196 amino acid peptide, with the start codon located in exon 1 and the stop codon in exon 3 (Fig. 28). The putative protein has a calculated molecular mass of 22.3 kDa and an isoelectric point of 9.26. Whereas no transmembrane domains are predicted, the analysis for specific motifs using the signature-recognition tools offered by InterPro[54] revealed that the first 26 amino acids probably function as a signal peptide. It also revealed that amino acids 99 to 109 and 138 to 148 present high similarity to the FARP (FMRFamide related peptide family) signature (Fig. 28). These two LPLRFGR motifs are characteristic of a large family of secreted neuropeptides (Price and Greenberg 1977 and Dockray et al. 1986). Due to the presence of this motif, the C7orf9 gene was named RFRP (Liu et al. 2001). It has been speculated that RFRP encodes a precursor protein for two (Liu et al. 2001) or possibly three (Hinuma et al. 2000) RFamide-related peptides (RFRPs), referred to as NPSF, NPVF and RFRP-1 to -3, respectively.

[54] http://www.ebi.ac.uk/interpro/

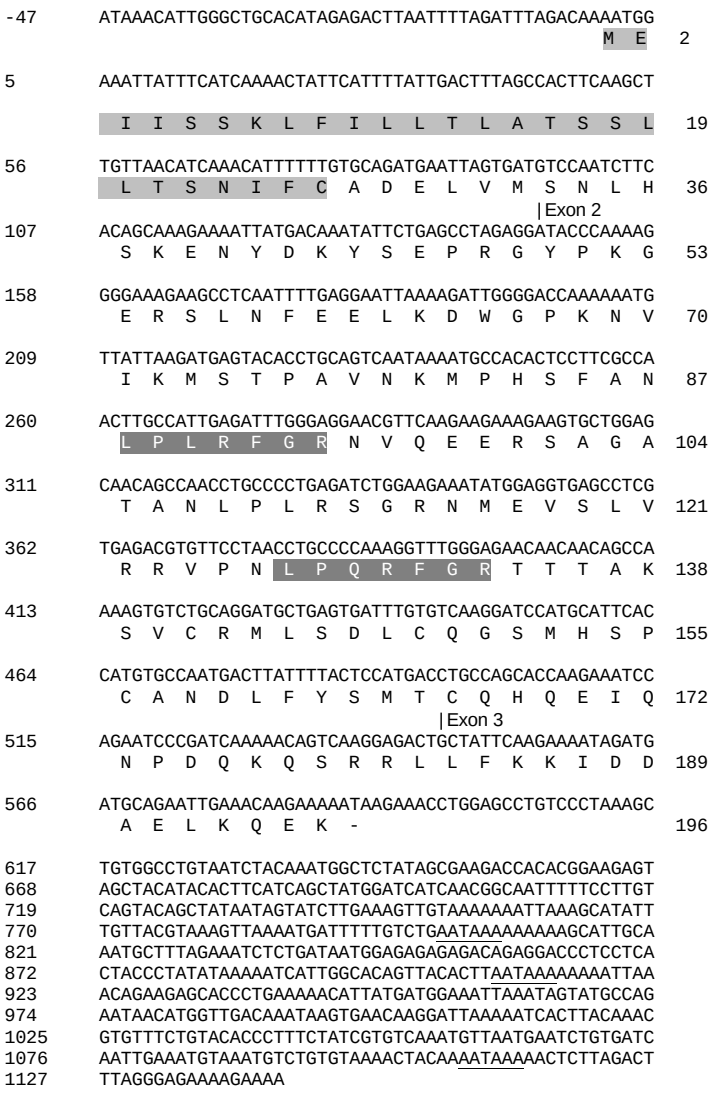


Fig. 28 Full-length cDNA sequence and putative protein of C7orf9

The nucleotide sequence is listed with the corresponding nucleotide numbering on the left. The amino acid sequence is reported underneath the cDNA sequence, with the amino acid listed beneath the first base of the codon (numbering on the right). A 26 amino acid signal peptide (shown by a light gray background) is predicted by the SignalP program. Two characteristic motifs for RFamides are depicted by a dark gray background. The three conserved poly-adenylation signals, AATAAA, are underlined.

To date the bovine, rat and murine orthologues of C7orf9 have been reported. A comparison of the proteins (Fig. 29) reveals that the homology between the human protein and the bovine, murine and rat proteins is of 72%, 60%, and 65%, respectively. While the human, bovine and murine proteins share the same amino- and carboxyl-terminal sequence, the rat protein has a longer carboxyl-terminal domain. An alignment of all sequences (Fig. 29) confirms high conservation of signal peptide and RFamide domains. In addition to the high homology of the orthologue proteins, the human sequence has a 49% identity with the Gonadotropin inhibitory hormone-related peptide 1 from the Japanese quail (GenBank acc. Q9DGD4).

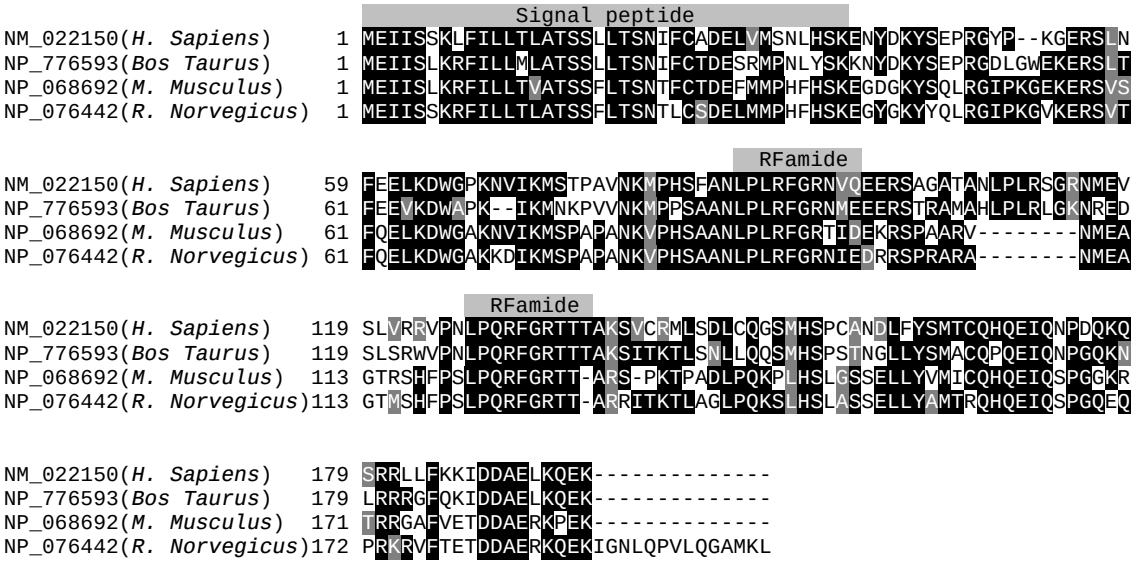


Fig. 29 Multiple sequence alignment of the C7orf9 orthologue proteins
The reference sequences of the human (*H. sapiens*), bovine (*B. Taurus*), murine (*M. musculus*), and rat (*R. norvegicus*) orthologue proteins were aligned using the using iterative pairwise method (Huang 1994). The majority of the amino acids are identical in all proteins (marked with black background). Amino acids which are similar chemical structure or properties are highlighted in grey. Since not all proteins have the same length and some have additional amino acids, gaps have been introduced by the alignment program and are shown as lines. As can be seen, the signal peptide and RFamide domains (detailed in the grey boxes above the sequences) are conserved in all species.

A phylogenetic analysis of the four proteins predicts the rat and mouse proteins to be more closely related as the human and bovine (Fig. 30).

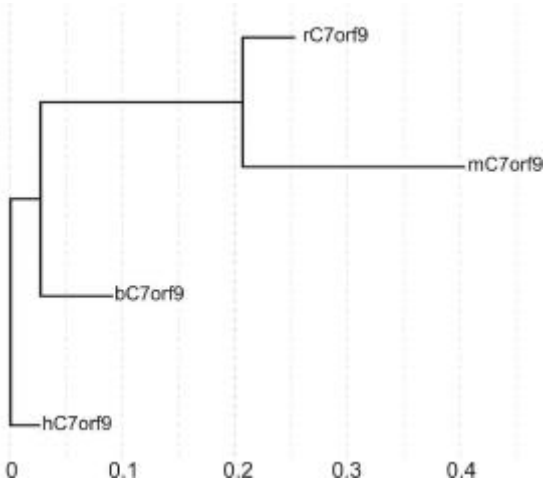


Fig. 30 Phylogenetic tree of proteins homologous to C7orf9
Sequence divergence between the C7orf9 proteins from *Rattus norvegicus* (r), *Mus musculus* (m), *Bos Taurus* (b), and *Homo sapiens* (h).

6.1.5. Polymorphisms in C7orf9

In the course of the cloning of C7orf9, three base exchanges were identified (Table 24). Two of them, a c.96G>C and a c.125A>G base change, result in non-conservative amino acid substitutions. Specifically, the changes are a methionine to isoleucine change at codon 32 and a substitution of aspartate at codon 42 to glycine, respectively. In contrast, the c.384C>T variant does not alter the amino acid sequence of the putative protein. To confirm the polymorphic nature of the changes, the frequency of each allele was determined using the most convenient method. The c.96G>C base change was investigated by SSCP analysis with primer pair A129E1F/A129E1R (Fig. 26 and Table

36) in 46 control individuals. The frequency of the minor allele was found to be 7% (Table 24). Even though the second base change (c.125A>G) is contained in the PCR product analyzed by SSCP, the evaluation of the shifts did not enable to distinguish between G homozygous and heterozygous individuals. This fragment was therefore sequenced from 13 controls and the polymorphic nature of the c.125A>G base change was confirmed since the minor allele was found in 27% of the Caucasian population (Table 24). The c.384C>T alteration was analyzed by restriction enzyme analysis of a 297 bp PCR fragment amplified from genomic DNA of 79 unrelated controls using primers A129F7 and A129R7 (Fig. 26 and Table 36). Digestion with *MnlI* resulted in two fragments if the C allele is present at position 431, whereas digestion of the T allele generated three fragments. The frequency for this allele was found to be of 27% for the minor T allele (Table 24).

Table 24 Single nucleotide polymorphism frequency for the two polymorphisms observed in the CYMD patients

Nucleotide exchange[a]	Location	Amino acid change	Allele frequency (minor allele)
c.96G>C	Exon 1	M32I	0.07 (n=92)
c.125A>G	Exon 1	D42G	0.27 (n=23)
c.384C>T	Exon 2	P128P	0.27 (n=158)

6.1.6. Analysis of C7orf9 as a candidate for dominant cystoid macular dystrophy (CYMD)

The retinal expression of C7orf9 as well as the localization of its genomic locus within the critical region for CYMD which is limited by D7S493 distally and D7S2444 proximally (Kremer et al. 1994 and personal communication), makes this gene an excellent candidate for the CYMD gene. Therefore, the DNA of two affected individuals (numbers 10084 and 4716) from a large Dutch CYMD pedigree used to establish linkage (Pinckers et al 1983) was analyzed. The complete coding region of the gene was investigated in both patients and an unrelated unaffected individual. The three exons were PCR-amplified using oligonucleotide primers A129E1F and A129E1R (exon 1), 129E2F and 129E2R (exon 2), and A129E3F and A129R6 (exon 3) (Fig. 26 and Table 36). Sequencing of all products revealed that patient number 10084 was heterozygous for the c.96G>C nucleotide substitution whereas individual number 4716 was heterozygous for the 384 bp SNP. Neither of these nucleotide changes is likely to be disease causing as both affected individuals belong to the same extended pedigree and would be expected to carry the same disease causing mutation. Moreover, the polymorphic nature of these base changes was confirmed (Table 24). Furthermore, heterozygosity of the two patients at the c.96G>C and c.384C>T sites confirms no large rearrangements such as a major deletion may have occurred. This fact rules out the possibility of a heterozygous deletion involving the entire gene, a situation that may otherwise have remained undetected by PCR amplification. To rule out intragenic rearrangements, the C7orf9 gene locus was further investigated in patient number 4716 and a control by Southern blot analysis. The genomic DNA was digested with restriction enzymes *EcoRI*, *HincII* and

PstI electrophoretically separated, blotted, and hybridized with a labeled probe obtained by amplification of retina cDNA with primer pair A129E1F/A129E3R (Fig. 26 and Table 36). Visual inspection of the band muster present in the control and the patient rule out any gross intragenic rearrangements (Fig. 31).

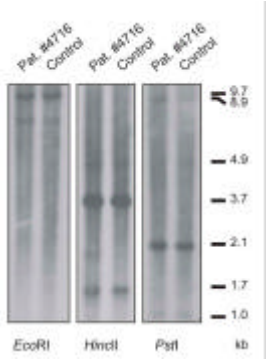


Fig. 31 Southern blot analysis of the C7orf9 locus
DNA from CYMD patient number 4716 and an unrelated control were digested with restriction enzymes EcoRI, HincII, and PstI and probed with a full-length transcript of C7orf9 amplified with primers A129E1F/A129E3R. All seven expected fragments are observed in both the patient and the control, thus no gross intragenic rearrangements are observed.

6.2. Cloning and characterization of C12orf7 (A038)

6.2.1. Assembly of the cDNA sequence of C12orf7

The initial sequences of C12orf7 were retrieved from cluster Hs.433492 which contained ten 5'- and 3'-ESTs which had been sequenced from five different clones. Four of the five clones were from the Soares retina library, the fifth was sequenced from the Soares placenta library.

Since the ends of the clones did not overlap, the 1.5 kb insert of clone zf50g06 was sequenced (Fig. 32). In order to identify additional clones which might extend the cDNA sequence, the Nathans retina library was screened. The screening probe, amplified from clone zf50g06 using primers A38F3/A38F (Fig. 32 and Table 36), was hybridized to approximately one million clones using duplicate filters. From the 24 positive signals observed after first round screening, the inserts of 11 clones (L-38-2-1, L-38-3-1a, L-38-4-1, L-38-7-1, L-38-11-1, L-38-12A-2, L-38-14-1, L-38-16, L-38-18A, and L-38-21-1) were isolated and sequenced. Analysis of the sequences determined that there was a variety of isoforms. Some of them, such as the sequence of L-38-4-1 did not seem to be spliced; others (e.g. L-38-7-1) spliced partially (Fig. 32). Comparison of the spliced sequences revealed that there were at least two different splice variants (e.g. L-38-18A and L-38-7-1) which differed by an insertion of 30 bp. Although both splice variants contained an ORF it was evident that sequences from the 5'-end were missing, because the 5'-ESTs from Hs.433492 extended further upstream. In order to characterize the 5'-end of the gene, three 5'-RACE experiments were undertaken. The first two rounds were done using the 5'-RACE System, version 2.0 from Life Technologies and the gene-specific primers A38F, A38R3, and A38R4 (Fig. 32 and Table 36). The amplified products prolonged the 5'-end by 106 bp. A third attempt using the RLM-RACE system and primers A38R3, A38R5 and A38R4 (Table 36) resulted in several products. Sequencing of two products revealed that the longest one (R-L6 5, Fig. 32) extended the 5'-end by 69 bp. The difference between the products was again due to an insertion, in this case of 27 bp.

In order to investigate the abundance of the different splice variants, retina cDNA was amplified with primers A38F2 and A38R3 (Fig. 32 and Table 36). This amplification resulted in the identification of additional unexpected splicing variants. Two of the products (P-A38 A8, P-A38 A9) were novel combinations of the already known alternatively spliced exons, but the third product (P-A38 A1) contained a novel splice form (Fig. 32). In order to investigate the proportion of transcripts retaining the last intron, a DNAse-treated retina cDNA was amplified with primers A38F5/A38R6 (Fig. 32 and Table 36). Interestingly, the IVS5-containing transcript had a double or threefold higher expression than the other isoform.

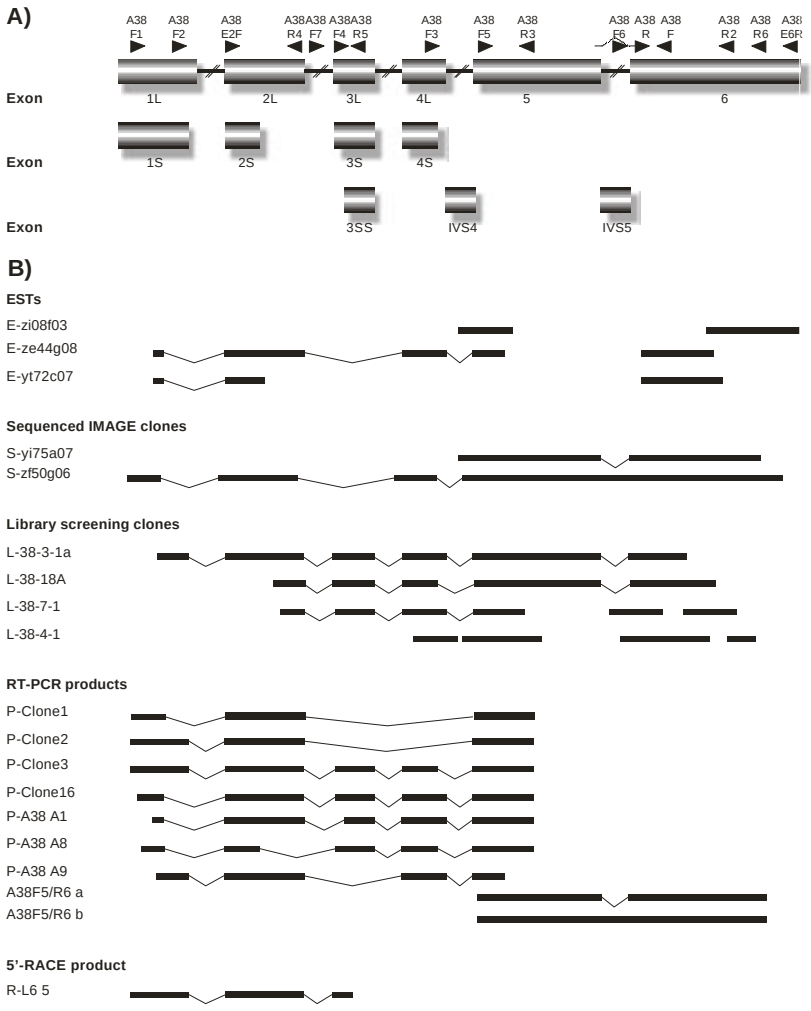


Fig. 32 Genomic organization of C12orf7 and selection of representative clones

A) Schematic organization of the C12orf7 gene which spans 3.6 kb. The exons are represented by the shaded boxes whose size is proportional to the correspondding exon size; the connecting lines (not proportional) represent the introns. Various transcripts which are assembled from alternatively spliced exons and may occasionally contain sequence from two intervening sequences (IVS4 and IVS5) have been found. Exons 1, 2, 3, and 4 present 2 or more alternative splice forms. They are identified for example as 3L, 3S, and 3SS for exon 3. The localization and orientation of the primers used in the cloning of the gene are indicated above the exons with the direction of the arrow indicating the orientation of the primer. In the case of primer A38R 14 bases anneal to exon 5 but the last four 3'-bases anneal to exon 6, this is represented by the broken line linking both halves.

B) The sequence of C12orf7 was determined by sequencing more than 40 clones, PCR-products, and 5'-RACE products. Only representative clones showing the different splice variants are depicted. In cases where the sequencing was not continuous a blank space between the lines can be observed (e.g. L38-4-1). Two different PCR products were obtained in a PCR amplification of retina cDNA with primers A38F5 and A38R6. They are indicated as a and b.

At this stage, the sequence of C12orf7 was deduced from the combination of sequences obtained from various clones, PCR-products, and RACE experiments. Although amplification of the various full-length transcripts in one PCR was tried repeatedly with different primer combinations none yielded results. We did succeed in amplifying the coding region of the gene in two overlapping fragments using primers A38F1/A38R3 and A38F5/A38R6 (Fig. 32 and Table 36). Surprisingly, the 5'-end PCR (A38F1/A38R3) amplified five different products from retina cDNA, whose sequences ranged from 580 to 871 bp (P-Clone1, P-Clone2, P-Clone3, P-Clone10, and P-Clone16). From these, two (P-Clone1, P-Clone2) were novel and result from the skipping of two exons (Fig. 32).

6.2.2. Genomic structure of C12orf7 and its isoforms

The longest cDNA sequence of C12orf7, excluding possibly unspliced transcripts, is 1709 bp long and is organized in six exons distributed along 3.6 kb of genomic sequence (Table 25). Alignment to the human genome assembly (April 2003) identifies the C12orf7 gene locus within chromosomal band 12q13.13. The chromosomal localization was verified by amplification of a 300 bp product only from chromosome 12 with primers A38F6 and A38F (Fig. 32 and Table 36) (Fig. 34). The sequence of each exon (Fig. 33) has been submitted separately and received a GenBank accession number which is listed in Table 25. To facilitate the identification of each exon isoforms, by convention the longest is identified by the exon number followed by the letter L, a short isoforms by the letter S, and the shortest by the letters SS. As already reported, there are multiple splice variants and isoforms. In some of them exons 3 and 4 are removed, whereas in others the intronic sequence between exons 4 and 5 or 5 and 6 may be retained. Since each isoforms presents a slightly different ORF it is difficult to evaluate the coding potential of the various isoforms. Regarding the 3'-end of the gene, two polyadenylation signals were identified. They are not the classical AATAAA signals, but the less common AGTAA and AAATAA. The first 'classical' polyadenylation signal is found 170 bp downstream of the last base of clone yi08f03, which contains the 3'-end of the gene. Another feature of the sequence is the 267 bp L1MC4a (LINE class) repeat localized within the 3'-end of the gene (Fig. 33).

Table 25 Exon / intron structure of C12orf7

Exon			3'-Acceptor Splice Site[a]		5'-Donor Splice Site[a]		Intron	
No.	GenBank acc. no.	Size (bp)	Sequence	Score[b]	Sequence	Score[b]	No.	Size (bp)
1L	AF517108	249			TCCgtgagt	5.84	1S? 2	236
1S	AF517109	222			CAAgtaaga	3.14	1L? 2	209
2L	AF517110	251	ccattctgtgttttagA	7.15	AGGgtgaga	2.68	2L? 3L	248
	AF517111						2L? 3S	252
							2L? 3SS	284
2S	AF517111	111	ccattctgtgttttagA	7.15	CGGgcaagg	7.51	2S? 3L	388
							2S? 3S	392
							2S? 3SS	424
3L	AF517112	130	gctgcccctccctcagA	4.71	CAGgtgtga	3.46	3? 4	175
3S	AF517113	126	cccctccctcagacagG	5.60	CAGgtgtga	3.46		
3SS	AF517114	94	ctaccacggcttccagA	7.63	CAGgtgtga	3.46		
4L	AF517115	141	ccttgctgccctgcagG	3.93	CAGgtgtgc	3.33	4L? 5	1071
4S	AF517116	111	ccttgctgccctgcagG	3.93	GCTgtgagt	4.62	4S? 5	1101
IVS4[c]	AF517117	1071						
5	AF517118	402	ctgtgctccttcccagG	3.27	TAGgtactg	6.91	5? 6	200
IVS5[d]	AF517119	200						
6	AF517120	536	ctcatcctcccaccagT	6.16				

[a] Exonic and intronic sequences in upper and lower case letters, respectively.
[b] Score of donor/acceptor splice site. According to published data (Berg and von Hippel 1998 and Penotti 1991) 99% of sites have a score of 0-11 (donor) or 0-20 (acceptor). Scores were calculated using the spreadsheet created by Christian Sauer, 2001.
[c] In some variants, the IVS is retained

The number and genomic organization of the various transcripts expressed from this gene is difficult to assess due to the large number of differentially spliced exons. Theoretically, up to 128 different isoforms may exist. This is calculated by multiplying the number of different exon combinations

(2x2x4x4x2x1). In the 5'-end of the gene (exons 1 thru 5) alone we have identified 11 different splice variants (Fig. 32). In the 3'-end of the gene (exons 5 and 6) two different splice forms are known (with and without IVS5). Therefore, it is likely that at least 22 isoforms of C12orf7 are expressed.

Exon 1L
GACCACCCGCCCCGCATGGGGCCCCARTCCCAGCTGCTTGATCCGGCTCAGCCCCGAGGTGTTTGCAGCAGCTCTTTATGAAAGTCCAGCCATCTGTTACCTGCGTT
GCTTCCTGGGGAGGGATAGTCCACCTGGAGGCATTCGGAGACCCAGTGATTGTGCTCCGYGGAGCCTGGGCTGTGCCCCGCGTTGACTGCCTCATAGATACCCTAC
GAACCCCAAGTAAGAAAAAACGACGACCCTCTCTCC

Exon 1S
GACCACCCGCCCCGCATGGGGCCCCAGTCCCAGCTGCTTGATCCGGCTCAGCCCCGAGGTGTTTGCAGCAGCTCTTTATGAAAGTCCAGCCATCTGTTACCTGCGTT
GCTTCCTGGGGAGGGATAGTCCACCTGGAGGCATTCGGAGACCCAGTGATTGTGCTCCGTGGAGCCTGGGCTGTGCCCCGCGTTGACTGCCTCATAGATACCCTAC
GAACCCCAA

Exon 2L
ATGCCAGCTGCATGAGAAAAGGGACTCACCTTCTGGTTCCCTGCCTGGAAGAGGAAGAGCTGGCATTGCACAGGAGACGGCTGGACATGTCTGAGGCACTGCCCTGC
CCGGGCAAGGAGACCCCCACCCCAGGCTGCAGGCTGGGGGCCCTGTATTGGGCCTGTGTCCACAATGATCCCACCCAGCTCCAAGCCATACTGGATGGTGGGGTCT
CCCCAGAGGAGGCCACCCAGGTGGACAGCAATGGGAGG

Exon 2S
ATGCCAGCTGCATGAGAAAAGGGACTCACCTTCTGGTTCCCTGCCTGGAAGAGGAAGAGCTGGCATTGCACAGGAGACGGCTGGACATGTCTGAGGCACTGCCCTGC
CCGG

Exon 3L
ACAGGCCTCATGGTCGCATGCWACCACGGCTTCCAGAGTGTTGTGGCCCTGCTCAGCCACTGTCCTTTCCTTGATGTGAACCAGCAGGACAAAGGAGGGGACACGG
CCCTCATGTTGGCTGCCCAAGCAG

Exon 3S
GCCTCATGGTCGCATGCTACCACGGCTTCCAGAGTGTTGTGGCCCTGCTCAGCCACTGTCCTTTCCTTGATGTGAACCAGCAGGACAAAGGAGGGGACACGGCCCTC
ATGTTGGCTGCCCAAGCAG

Exon 3SS
AGTGTTGTGGCCCTGCTCAGCCACTGTCCTTTCCTTGATGTGAACCAGCAGGACAAAGGAGGGGACACGGCCCTCATGTTGGCTGCCCAAGCAG

Exon 4L
GCCACGTGCCTCTAGTGAGTCTCCTGCTCAACTACTATGTGGGCCTGGACCTGGAACGCCGGGACCAGCGGGGGCTCACGGCGTTAATGAAGGCTGCCATGCGGAA
CCGCTGTGAGTGCGTGGCCACCCTCCTCATGGCAG

Exon 4S
GCCACGTGCCTCTAGTGAGTCTCCTGCTCAACTACTATGTGGGCCTGGACCTGGAACGCCGGGACCAGCGGGGGCTCACGGCGTTAATGAAGGCTGCCATGCGGAA
CCGCT

IVS4
GTGTGCGGGGCCTGGACCGGGGTGTGTGGCCTCCAGTCCCTCCTCCAAGCCTTCCCACCAGACACTAAGTCAGCTGTGATTATTTGCAGAAAGGAGAGAGGTGGAGA
TGGGGGATCAATCTAGTTCACCTTCCTGGAGGGGGGGGCACATGATTTGGGCTTCGTACAATCAACCAAGTCCTGCTATTGATAGCTCAGACATTGTGTGTGGAGGTC
CCAGGAAGGAAAGTGTGGGAGGGGAACCTAGCCCAGCTGGAAAATCTAACACGAAAAAAGGCTGGCATCTGGAACCAGCCCCTGTTCCTAGCTGAGTGGCTTCCCTC
TTTGGGCCTCACCAGTATTGTGAGGGTGTTAGCTTTCCCTGGGACTACCTCCAACTCTGACAAGCCAGGCTTCCAGGGGATCCAGGGAAGTGTCGTTGAGGCCTGTG
GCTCTGTGGGGTCGTTCAGGGGGAGGTGCGGTGAAGTCCACATTGTCCATGGAGTTGTTTGGGGGCCCTTCTTCCCCGCAGTGGGGCTGCCCTCTGCTGGTCACTCT
GGGGGACCCCTGCCTCCATTTTTCCCTTCCCCACACACCCCACTGGGTTTGGAGTGAAAAAGGAGTGAATGAAAAAGGAGGGCGCTCTACACCCTTTCTTTTGTCTTA
GAGTGACTGCTCCTCCCACAACCCCAAGATGGGAGAGGGAGATGGTGAGAAGACAAAATCACCAACCCCTATCCCGCCCCATGTCACCCCCTGTGCTCCTTCCCAG

Exon 5
GTGCTGACCTGACAGCAGTGGACCCTGTTCGGGGCAAGACGGCCCTGGAATGGGCAGTGCTGACCGACAGCTTCGACACCGTGTGGAGGATTCGGCAGCTGCTGAG
GCGGCCCCAAGTGGAGCAGCTTAGCCRGCACTACAAGCCCGAGTGGCCGGCCTTGTCCGGGCTCGTGGCCCAGGCCCAGGCCCAGGCCCAGGTTGCCCCTTCACT
CCTAGAACGGCTGCAGGCTACCTTGAGCCTCCCCTTTGCCCCGTCTCCTCAGGAGGGGGGTGTTCTGGACCACCTTGTGACTGCCACAACCAGCCTGGCCAGTCCCT
TCGTCACCACTGCCTGCCACACTCTGTGCCCTGACCATCCACCTTCGCTGGGCACCCGAAGCAAGTCCGTGCCAGAGCTGTTAG

IVS5
GTACTGCCCCGCCCCCTCCCCTGGTTCCCCAGTCCCCGCCAGGGAGTCCCCAGAGGTCCCCGTGGGTCTTCGTCCCCTACCAGAGCCCTCAGGGCATATTGAGCAA
GTGCCTTCAGTGGCTACAACCCAGGGATAGCACCAGCCCCAGGCCCCAAGTCCCCAAGATCCTCCTCTCCAAGGCATCCTCATCCTCCCACCAG

Exon 6
TGCCAGCCGAAGCCCAGTCCTTCAGGACACCAAAGTCTGGCCCTTCCTCTCTGGCGATACCAGGAGCTCAGGATAGAGAAGARGAAACAGGAGGAGGAGGCCAGAATGG
CACAGAAGTAGRGGAAGATGGGATAGGACAGGCTGGGAACAGGTAATCAGGCCCCTCCCAGGGCTTCTTTCCCCTCTGGAGTGCCTCCGGCCTCCCCATCCACCTCTGCC
TAAGTAAATCTGCTCTCAACCTATATATATACAAGGTCATTCATTCTAGCATTGTTTGCAAGAGTGAAAGAGTGGAAACACCCGAAGTGTCCATCAGTAAGGGACAGGCT
AGATTGATTACGGATGTAATTGCTGTCCATCCATACAGAGCATACTCTACAGTGTATTCTAAATAAGACTAAGGAAGCTGTTTATATTCTGATATGAAACTACCATCAAG
ATGTATAAAGTAAAAATAACTAAGGAGTGGAACAGTGTATATGGCATATTATTATTTGTCGAAGTAAAATTTTCACGAAGATAAACATGACTAAGA

Fig. 33 Exon sequences of C12orf7
Six different exons have been identified, of which most have alternative splicing donor and acceptor sites. The IVS4 and IVS5 sequences are also shown, as they are retained in a number of transcripts. The sequences which differentiate the different splice variant are identified by the grey background. In the case of exon 3, variant 3S is missing the first four bases (light grey) and variant 3SS is missing the 36 bp identified by the light and dark grey backgrounds. The black background indicates the fragment of the sequence which is part of a LINE repeat. The red-highlighted bases indicate the position of bases whose allele frequency has been investigated (see Table 26). Green-colored letters indicate possible polyadenylation signals. The cDNA sequence of each exon has been submitted to GenBank and can be retrieved under acc. no. AF517108 thru AF517120.

6.2.3. Expression analysis of C12orf7

The expression profile of C12orf7 was investigated by RT-PCR. The first RT-PCR, done with primers A38F and A38R (Fig. 34 and Table 36) determined that the gene is expressed specifically in retina.

Amplification with A38F4/A38R3 (Fig. 34 and Table 36) confirmed this result by revealing a strong PCR product in retina, but also very weak products in brain, fetal brain, liver, trachea, spleen, testis, and salivary gland (Fig. 34).

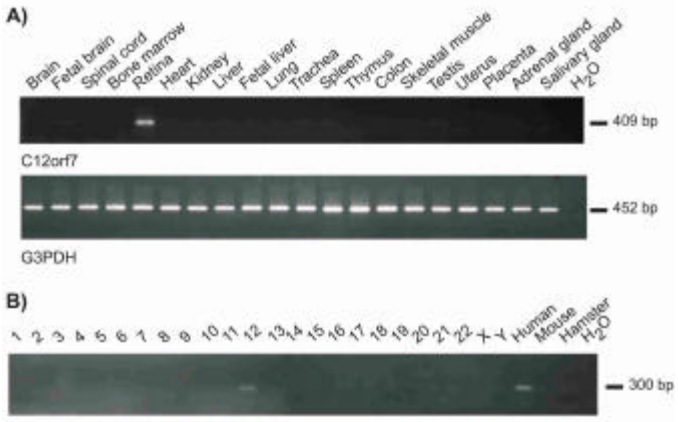


Fig. 34 Expression analysis of C12orf7 and verification of the chromosomal localization
A) C12orf7 is expressed at high levels only in the retina as shown by amplification with primers A38F4/R3 in a panel of 20 human tissues. Weak expression could be observed in brain, fetal brain, liver, trachea, spleen, testis, and salivary gland. As a control of the quality and quantity of the cDNAs used, an amplification of the cDNAs with G3PDH is shown below.
B) The chromosomal localization of C12orf7 was verified by amplification with primers A38F6/A38F resulting in a 300 bp product in the hybrid cell line containing human chromosome 12 only.

Northern blot analysis not only reveals the expression of a transcript but also indicates the size of the transcript. For C12orf7, knowing the transcript size would be helpful to possibly establish a major transcript sizes. Towards this end six different membranes (Lab ID A, D, H, O, V, and SFIp96) were hybridized at different time points with three different probes amplified with primers A38F/A38R, 38F4/A38R, and 38F2/A38F, respectively. Even though various membranes and probes were used, none of the hybridizations identified a specific hybridization signal. Possibly, the Northern blot technique does not reach the level of sensitivity to detect the large number of isoforms which each may be expressed at low levels.

6.2.4. Identification of C12orf7 orthologues

Using the tools available at the Golden Path server[55], we identified the orthologous genes of mouse and rat C12orf7. The probable C12orf7 rat orthologue is encoded on chromosome 7. The rat sequence is based on gene predictions and three ESTs (found in Rn.42318) sequenced from a rat eye library.

The mouse orthologue is located on chromosome 15 and is conserved in a syntenic block. A number of transcripts, all of retinal origin, provide sufficient evidence that the gene is also expressed in mouse. These sequences are grouped in UniGene cluster Mm.152952 which contains 20 ESTs. A comparison of the human and mouse sequences determined that exon 1 is not conserved between the two species. In mouse, only one splice form of exon 2, homologous to human exon 2L is evident. In many murine transcripts exon 2 is spliced out, something that was not seen in any human transcript. As seen in human, there are also different splice variants of exon 3 in the mouse. The two murine exon 3 variants also differ by four bases. Therefore, human exon 3L is homologous to mouse exon 3L and human 3S is similar to mouse 3S. None of the reported murine sequences encode the shortest

[55] http://genome.ucsc.edu/cgi-bin/hgBlat?command=start

version of exon 3 (human 3SS). The mouse gene has two versions of exon 4; human 4L is homologous to mouse 4S. In humans, no sequence has been observed to contain an exon 4 variant homologous to the longer mouse exon 4 (4L). Exon five of the mouse is 1508 bp long and may be spliced out (e.g. in AK044543). Since the homology between the human and mouse sequences decreases in exon 5, it is not possible to determine the 3'-end of the murine C12orf7 gene.

6.2.5. *In-silico* analyses of the putative C12orf7 protein

The existence of multiple transcripts of C12orf7 implies that a number of distinct proteins may be encoded by this gene. Many isoforms contain more than one ORF, further increasing the complexity at the protein level.

Exon 1L or 1S, which differ by 27 bp, have the same ORF. A transcript containing exon 1S will be just 9 aa shorter. In the case of exons 2L and 2S, the insertion of 140 bp is not a multiple of three, and thus the protein encoded by each exon variant would be different. If, however exons 2S and 3S are spliced together, the protein is very similar to the one translated from a transcript with exons 2L and 3L. The usage of exon 3L or 3SS results in a loss or gain of 12 aa. Similarly, exons 4L and 4S differ by 30 bp resulting in a peptide difference of 10 aa between 4L and 4S.

Based on the known partial-sequence transcripts, 20 hypothetical isoforms of C12orf7 were assembled and translated. Since in some cases the longest ORF may not be the one used for translation *in-vivo*, all possible ORFs were included and the total of hypothetical proteins ascended to 25. The encoded ORFs ranged in size from 135 to 471 aa and are characterized by a common sequence. Some are shorter, others are missing the N-terminal, and another group has deletions of internal amino acids. In general, they all encode an identical core protein.

The longest ORF is encoded by isoform C12orf7_v1 which is composed of exons 1L, 2L, 3L, 4L, 5, IVS5, and exon 6. The 1909 bp transcript encodes a 471 aa protein. The second (C12orf7_v2) and third (C12orf7_v3) longest ORFs are the ones from transcripts translated from exons 1L, 2L, 3L, 4L, 5, and 6 (416 aa) and 1L, 2L, 3L, 4S, 5, and 6 (406aa). The inclusion of IVS5 in a transcript results in an alternative carboxyl-terminal, but does not alter any of the presently-known functional domains (Fig. 35). Apart from the differences in molecular weight (55.1, 44.7, and 43.7 kDa), the inclusion of the last intronic sequence has a great effect on the isoelectrical point which is of 8.68 for C12orf7_v1 compared to 6.19 and 6.32 for C12orf7_v2 and C12orf7_v3, respectively.

Ab initio prediction of protein function by ProtFun[56], assigns the proteins of C12orf7_v1 and _v2 a cellular role related to the cell envelope, whereas the C12orf7_v3 protein would be involved in energy metabolism. Additional evidence supporting a role in the cell envelope comes from the identification of five ankyrin repeats in C12orf7_v1 and _v2 and four in C12orf7_v3 (Fig. 35). Aside from a nuclear localization signal present in all three isoforms, no additional signals or motifs were found for any of

[56] http://www.cbs.dtu.dk/services/ProtFun/

the variants. The transmembrane prediction programs report ambiguous results as to the existence of such domains. The TMHMM[57] algorithm does not predict transmembrane domains, whereas TMPRED[58] identifies four helices with the C- and the N-termini of the protein located inside the cell.

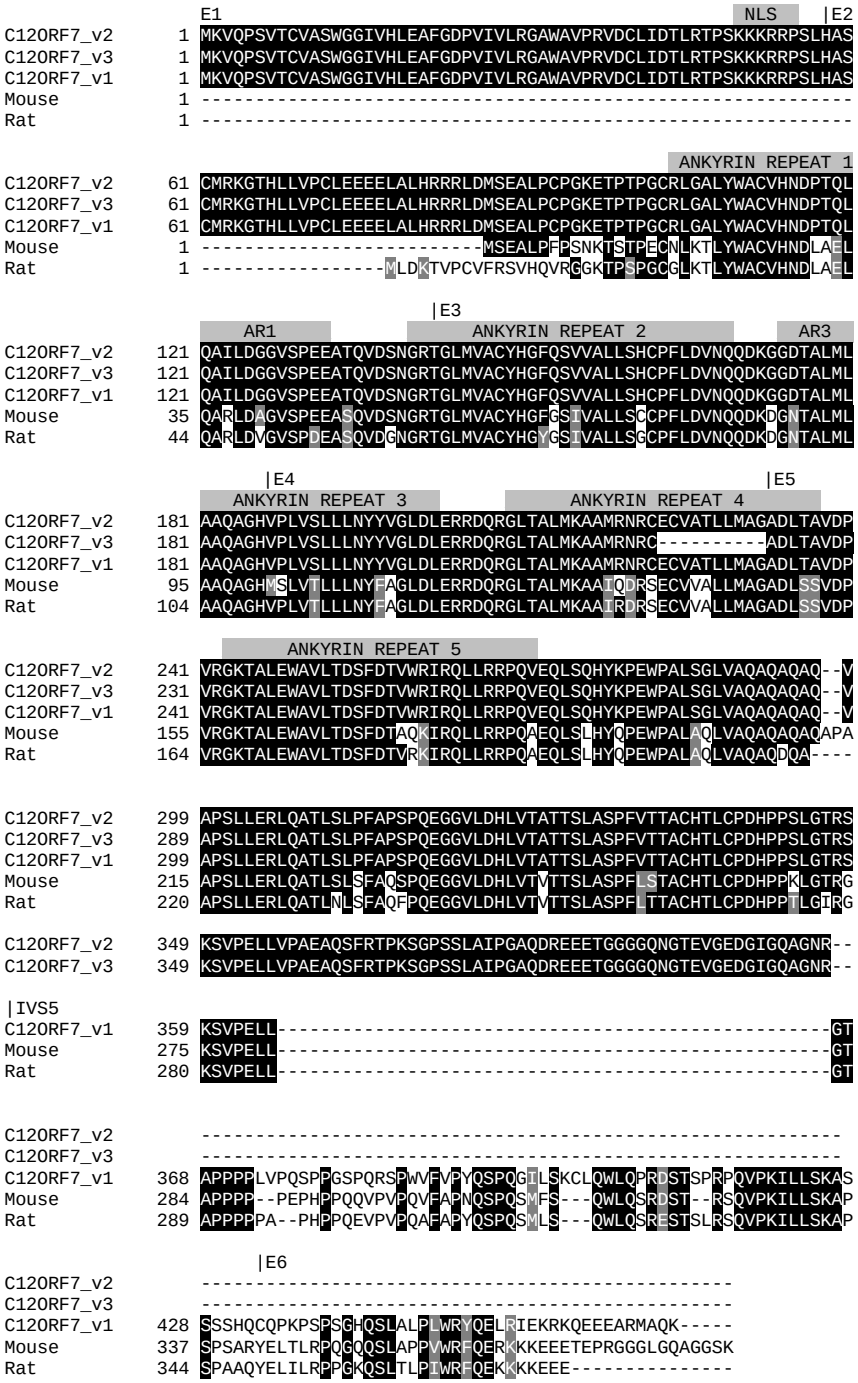


Fig. 35 C12orf7 putative proteins and comparison with the mouse putative protein
The longest possible ORF of C12orf7 is 471 aa long and is encoded by exons 1L, 2L, 3L, 4L, 5, IVS5, and 6 (C12orf7_v1). C12orf7_v2 and _v3 are encoded by transcripts which contain the same exons but without the IVS5 (C12orf7_v2) and with the short version of exon 4 (C12orf7_v3). The mouse protein (Mouse) was translated from a hypothetical assembly based on clones BM936148 and BU504698. It is homologous to exons 1, 2, 3L, 4S, 5, 5IVS, and 6. The rat sequence corresponds to the hypothetical prediction by GenomeScan XP_235671. The beginning of each human exon is indicated by a vertical line followed by the exon number. As can be seen, the retention of the IVS5 leads to an alternative C-terminal, but does not affect the ankyrin repeats (indicated by a grey box above the alignment). Note that the empty spaces in the C12orf7_v1, mouse, and rat sequences do not indicate a gap, but have been introduced to facilitate the visualization of the alternative carboxyl ends. The deletion of 10 aa in the C12orf7_v3 leads to the loss of ankyrin repeat 4. All three human putative proteins contain a nuclear localization signal (NLS).

6.2.6. Analysis of related proteins

To learn more about the C12orf7 proteins the non-redundant protein database was queried with two fragments, the N- and C-terminal regions of the proteins; the ankyrin repeats were removed. The search identified the putative orthologue rat protein XP_235671 and mouse BAC31973. The mouse protein is found in chromosome 15 while the rat is located in chromosome 7. An alignment of the three

[57] http://www.cbs.dtu.dk/services/TMHMM-2.0/
[58] http://www.ch.embnet.org/software/TMPRED_form.html

human isoforms with the mouse and rat hypothetical proteins reveals a high sequence homology from amino acid 77 to 456 of C12orf7_v1 (Fig. 35). Another finding was the human predicted gene XP_293937 (LOC345711) which maps to chromosome 5p15.31. The C-terminal part of C12orf7, beginning after the last ankyrin repeat, has an identity of 32% with the C-terminal part of the LOC345711 protein. A number of short hits with varied African swine fever virus proteins were also found.

As already described there are at least four mouse isoforms. For the comparisons the same exons used for the C12orf7_v1 human protein were assembled. The resulting 2061 bp mouse mRNA encodes a 385 aa protein with a molecular weight of 41.2 kDa and an isoelectrical point of 6.26. ProtFun[59] predicts this protein to belong to the voltage-gated ion channel group of the ontology classification and to have an enzymatic function in the cell envelope. Analogous to the C12orf7_v1 protein, to which it is most similar, it has five ankyrin repeats. It does not contain a nuclear localization signal, and therefore is predicted to be localized in the cytoplasma with a probability of 48%. The similarity between the mouse and human C12orf7_v1 proteins, both of which retain the last intron, is 73% in the region that is shared. Since the similarity is high also in the translated intronic region, it appears that this isoform may be real.

6.2.7. Identification and characterization of single nucleotide polymorphisms contained in C12orf7

Analysis of the sequences originating from C12orf7 revealed the existence of 19 different sequence variations including base changes, duplications, and deletions. A description of each variant, the allele frequencies, and the method used to determine allele frequencies is listed in
Table 26. The frequencies of four of the probable base changes were determined in a large cohort of control chromosomes. The frequency of c.904A>G was determined in 48 subjects by amplification of the locus with primers A38F8/A38R2 and subsequent digestion with *MspI*. The major allele, A, had a frequency of 0.74. The frequency of the c.523A>T polymorphism was estimated by amplifying genomic DNA with primers A38F7 and A38R5 (Fig. 32 and Table 36) and subsequent sequencing. The frequency was determined to be 0.61 for the major allele A. The frequencies of c.1256G>A and c.1294G>A was determined by SSCP analysis of the PCR product amplified with primers A38F6 and A38F (Fig. 32 and Table 36).

Table 26 Sequence changes found in the C12orf7 sequence

Description[a]	dbSNP ID	Allele frequency (minor allele)	No. of analyzed chromosomes	Method

[59] http://www.cbs.dtu.dk/services/ProtFun/

c.27A>G	rs697633	0.42	1276	Invader assay
c.119A>G	-	n.d.	-	
c.167T>C	rs697634	0.48	1484	Invader assay
c.287T>C	-	n.d.	-	
c.438C>T	-	n.d.	-	
c.523A>T	-	0.39	84	Sequencing
c.904A>G	-	0.26	96	Enzymatic digestion
c.962_967dupGGCCCA	-	n.d.	-	
c.1072C>A	-	n.d.	-	
c.1110A>G	-	n.d.	-	
c.1256G>A	-	0.01	90	SSCP
c.1287G>T	-	n.d.	-	
c.1290G>A	-	n.d.	-	
c.1294G>A	-	0.01	90	SSCP
c.1306T>G	-	n.d.	-	
c.1307A>C	-	n.d.	-	
c.1312A>C	-	n.d.	-	
c.1392delC	-	n.d.	-	
c.1422_1423dupTA	-	n.d.	-	

[a] As there are many different ORF, base number one of the cDNA sequence is considered to be the first base of the transcript (Fig. 33).
n.d. Not determined

6.3. Cloning and characterization of three novel isoforms of the metabotropic glutamate receptor 7 (GRM7)

6.3.1. Cloning and genomic organization of the GRM7 isoforms

In the course of the UniGene project, cluster Hs.59956 was identified as being neuronal-specific since primers A48F and A48R (Fig. 36A and Table 36) amplified a product only from retina and cerebellum cDNAs. Alignment of the four clones contained at that time in cluster Hs.59956 to Homo sapiens chromosome 3 clone RP11-329A2 (GenBank acc. no. AC077690) revealed that the 5'-ends contained no open reading frame (ORF) and were not spliced, whereas the 3'-ends spliced in two exons. Detailed bioinformatic analyses of genomic clone RP11-329A2 done by programs contained in the NIX analysis package[60] revealed that the 3'-ends of the ESTs mapped to the last intron of the glutamate receptor, metabotropic 7 (GRM7) gene (Fig. 36A). The results of similarity searches in the EST and non-redundant databases suggested that the Hs.59956 sequences could represent novel alternative exons of GRM7. This hypothesis received support from a report of a variant of GRM7, GRM7_v2 (Flor et al. 1997). This variant, referred in the original publication as mGlu7b (GenBank acc. no. A43639) contained an additional 92 bp exon located in the last intervening sequence (IVS) of GRM7. Further evidence came from human clone AW511101 which overlapped not only with exon nine of GRM7 (NM_000844) but also with the 3'-ESTs of Hs.59956 (Fig. 36). Alignment of the AW511101 cDNA sequence to the genomic clone defined additional putative exons, which were identified as 9b-e (Fig. 36).

[60] http://www.hgmp.mrc.ac.uk/Registered/Webapp/nix/

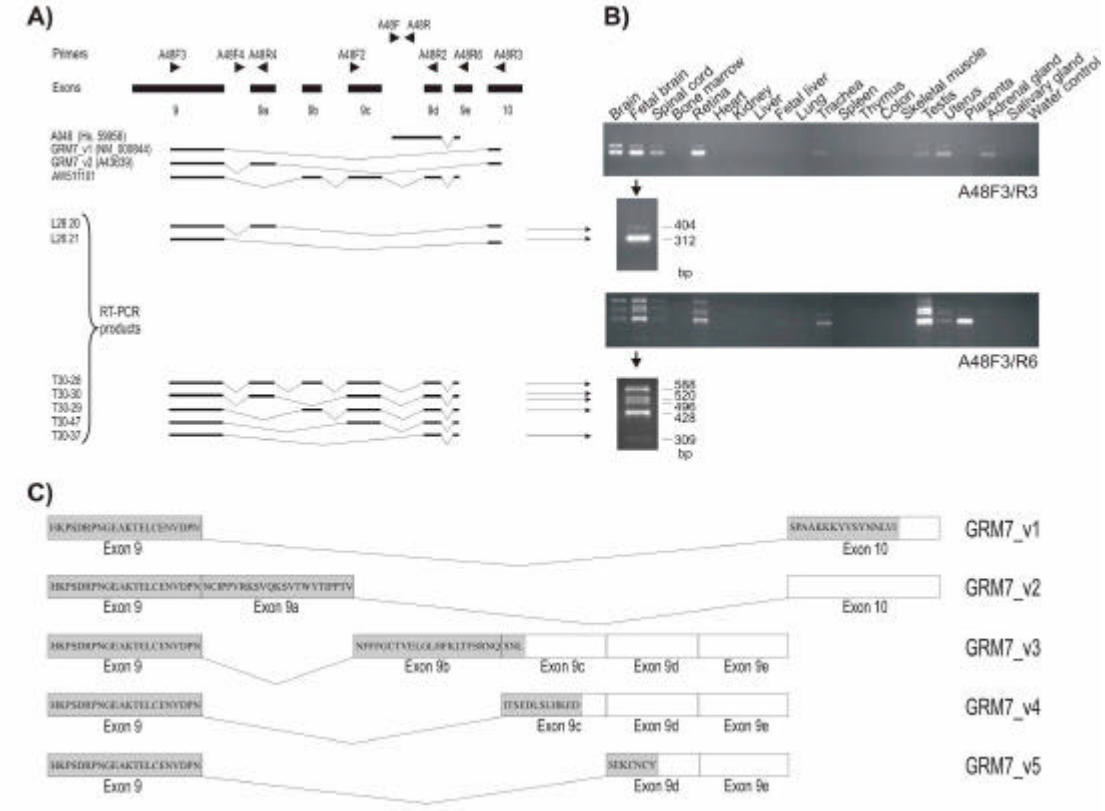


Fig. 36 Schematic representation, expression profiling, and putative protein sequences of the distinct C-termini of the GRM7 isoforms

A) Diagrammatic representation of the gene structure encoding the C-terminal region of GRM7. Oligonucleotide primers used in the analysis of RT-PCR products are given above the respective exons (arrowheads). RT-PCR fragments corresponding to amplified products in (b) (arrows) were sequenced and their genomic organization determined.

B) Expression analysis of GRM7 subtypes in a panel of 20 human tissues. Primer pairs used and sizes of PCR products are indicated.

C) Translation of exon 9 of GRM7 and alternatively spliced exons 9a-e and exon 10 reveals five distinct C-termini. Open boxes represent non-coding sequences.

To investigate if the exons found in the last IVS of GRM7 could be novel alternative splice forms of the gene, primers A48F3 and A48R3 (Fig. 36 and Table 36) were designed to align to exon 9 and 10 of GRM7. The PCR amplification consistently resulted in only two products of 404 and 312 bp (Fig. 36B) which were cloned. Sequencing of clones L26 20 and L26 21 revealed that the 312bp fragment represents the 3'-end of the known GRM7_v1 isoform while the 404 bp product corresponds to the previously reported alternative GRM7_v2 transcript (Fig. 36).

To assess whether Hs.59956 may constitute a novel isoform of GRM7 we analyzed 1 µl aliquots of six different retina cDNA libraries (DKFZ1, DKFZ2, DKFZ3, DKFZ4, C1F1 and C1F2) with primers A48F2/A48R2 (Fig. 36 and Table 36). No products could be amplified from any of the libraries.

Amplification with a primer designed to anneal to exon 9 of the GRM7 gene (A48F3) and another located in the last exon of AW511101 (A48R6) resulted in five products of 588, 520, 496, 428, and 309 bp in neuronal cDNAs (Fig. 36B). The products from fetal brain were cloned and representative clones of each product, labeled T30-28, T30-30, T30-29, T30-47, and T30-37, were sequenced (Fig. 36A). The sequence of clone T30-8, which contained the biggest insert, spliced in six exons when compared

with the genomic sequence (Fig. 36). Since the first exon corresponded to exon 9 of the GRM7 gene, the additional exons were identified as 9a-e (Table 27). Thus, with the A48F3/A48R6 PCR it was possible to confirm the existence of alternative transcripts of GRM7 which do not contain exon 10. The 520 bp product contained in clone T30-30 lacked exon 9b, the 496 bp fragment (T30-29) exon 9a, the 428 bp fragment (T30-47) exons 9a and 9b, and the 309 bp product (T30-37) exons 9a–c (Fig. 36A, Table 27). A thymine (T) stretch of variable length was identified in exon 9b. Whereas the genomic sequence NT_005927 contains 11 Ts, clones T30-28, T30-21, AW511101, and CA313509 had 10 Ts, and clone T30-29 contained 9 Ts. The number of thymines in this stretch was determined by amplification of genomic DNA from four subjects (AMD89, 90, 91, and 92) with primer pair A48F4/A48R4 (Fig. 36 and Table 36) and sequenced using Beckmann-Coulter technology. All four sequences contained 10 thymines in the stretch. Therefore, for the construction of the ORF of the novel isoforms, a stretch of 10 Ts was assumed.

Table 27 Exon / intron structure of GRM7[c]

Exon		3'-Acceptor Splice Site[a]		5'-Donor Splice Site[a]		Intron	
No.	**Size (bp)**	**Sequence**	**Score[b]**	**Sequence**	**Score[b]**	**No.**	**Size (bp)**
1[d]	668			CAGgtaggg	2.70	1	284,544
2	217	tttcttgtctttgcagA	3.01	CAGgtagga	2.95	2	152,015
3	142	ttcttctcttaaacagG	4.86	AAAgtaaga	3.36	3	7672
4	155	atatttctttccacagG	2.74	AAGgtatgg	3.35	4	108,370
5	141	atttaactctgtttagG	9.02	CAGgtaatt	2.89	5	37,443
6	201	ttgttttaatgtgcagG	6.60	ATGgtgagt	1.78	6	8775
7	140	tttctgtcttccttagG	3.69	AATgtgagt	2.54	7	116,699
8	936	atttcttatgttacagA	5.41	AAGgtaagt	0.22	8	100,691
9	247	gtgttgtgtctcctagC	7.64	ACAgtaagt	3.74	9? 9a	4590
						9? 10	60,061
9a	92	tttacttttcctgtagA	4.34	AGGgtaaga	2.60	9a	4590
9b	68	tatctcaactttgcagA	6.99	AAGgtaatc	4.04	9b	103
9c	119	ctgccctttttcatagT	5.25	TCTgtaagt	4.99	9c	2304
9d	60	tgttgtttttattcagG	4.69	GAGgcaagt	6.25	9d	670
9e	66[e]	tttctgtttcctctagG	3.14				
10	1155	ctaatttttctttcagG	3.30				

[a] Exonic and intronic sequences in upper and lower case letters, respectively.
[b] Score of donor/acceptor splice site. According to published data (Berg and von Hippel 1998 and Penotti 1991) 99% of sites have a score of 0-11 (donor) or 0-20 (acceptor). Scores were calculated using the spreadsheet created by Christian Sauer, 2001.
[c] The organization was based on reference sequences NM_000844, AF458052, AF458053, and AF458504.
[d] The sequence reported as exon 1 begins at base pair 303 of the reference sequence NM_000844. The exact location and organization of the first 302 bp has not yet been accurately determined.
[e] Partial length; the exon has not been cloned in its entirety.

On the basis of the sequenced PCR products it was possible to verify the existence of two known GRM7 variants and identify five new isoforms (Fig. 36) which do not contain exon 10 in their 3'-end. The sequences of the novel splice variants, transcribed from chromosome 3p26.1, have been submitted GenBank and have been deposited as GRM7_v3 (AF458052), GRM7_v4 (AF458053), and GRM7_v5 (AF458054). The latter variants all contain exon 9e in their 3'-end, but since the UTR has not been investigated, it could be possible that they contain yet another 3'-end exon. Two of the newly identified transcripts (containing exons 9a-9e and 9a + 9c-e) have not been submitted as novel splice variants as the putative protein they encode are identical to GRM7_v2 and it is not clear what the functional consequences could be. It should be noted that the exon numbering of GRM7 may be

subjected to changes because no publication has yet reported the organization of the 5'-end of the gene and the first 302 bp do not align to the current chromosome 3 sequence.

6.3.2. Expression analysis of the GRM7 isoforms

In order to assess the expression pattern and abundance of the GRM7 isoforms, two semi-quantitative PCR amplifications were done in a panel of 20 human first-strand cDNAs from different tissues (Fig. 36B). Both PCRs were done with the same forward primer (A48F3) but different reverse primer. The use of reverse primers A48R3, which primes in the 3'-UTR of the GRM7_v1 and _v2, and A48R2, which primes in the 3'-UTR of the novel GRM7 isoforms, allowed the amplification of all GRM7 isoforms in two PCRs (Fig. 36). The amplification with primers A48F3/A48R3 which amplify variants 1 and 2 showed that both isoforms, although with a preference towards GRM7_v1, are expressed in retina, neuronal tissue, and to a minor extent, in trachea, testis, uterus and adrenal gland (Fig. 36B). Expression of variants 3, 4, and 5, was determined with primers A48F3 and A48R6 (Fig. 36 and Table 36). In the neuronal tissues (brain, fetal brain and spinal cord) all five transcripts are found. In retina, four of the variants are expressed; while the 520 bp product, which contains exons 9, 9a, and 9c-e, is absent. It is noteworthy that whereas in brain the most common isoform is the one represented by the 588 bp product, in retina the most abundant isoforms are GRM7_v3 and GRM7_v4. In trachea and placenta only the GRM7_v4 is present while in testis and uterus the 520 bp product and GRM7_v3 and GRM7_v4 are exclusively expressed (Fig. 36B).

6.3.3. *In-silico* analysis of the putative GRM7 isoforms

The already reported protein isoforms GRM7_v1 and _v2 have distinct C-termini and a length of 915 and 922 aa, respectively. GRM7_v1 has a unique 16 aa terminus encoded by exon 10 which is replaced by 23 aa encoded by exon 9a in GRM7_v2. The alternative exon usage in the 3'-end of the novel isoforms also leads to distinct carboxyl-terminal domains starting at amino acid 900 (Fig. 36C). In-frame translations of the 588 and 520 bp products determined that their putative proteins that are identical to GRM7_v2, whereas the 496, 428 and the 309 bp fragments encode the novel isoforms GRM7_v3 (overall length of putative protein is 924 aa), _v4 (911 aa) and _v5 (906 aa), respectively (Fig. 36C). In GRM7_v3 the 25 unique amino acids are encoded by exons 9b and 9c, in GRM7_v4 exon 9c encodes for the 12 aa, and in GRM7_v5 the 7 aa are encoded by exon 9d (Fig. 36C). Thus, GRM7_v3 encodes the longest protein followed by GRM7_v2, v1, v4, and v5.

Protein and motif databases were searched for known domains. No significant homology was found between the putative carboxyl-terminal tails of the novel isoforms and the other GRM family members. Likewise, searches against the Swissprot[61] and non-redundant databases[62] failed to identify sequence identities to known genes or proteins. The only novel features found in the variants include ASN glycosylation sites at amino acids 920–923 in GRM7_v3 and at amino acids 899–902 in GRM7_v4.

[61] http://www.ncbi.nlm.nih.gov/BLAST/
[62] http://www.ncbi.nlm.nih.gov/BLAST/

Unique for GRM7_v4 is the prediction of a CK2 phosphorylation site extending from 901–904 aa. As for GRM7_v5, a putative protein kinase C phosphorylation is present at amino acid residues 900–902.

6.4. Cloning and characterization of C1orf32 (A166)

6.4.1. Assembly of the C1orf32 cDNA sequence

The interest to clone this gene stemmed from the finding of two UniGene clusters expressed preferentially in neuronal tissues that mapped only 20 kb apart from each other. Hs.129907 (Lab ID: A65) was a cluster with seven clones. Three of them derived from retina (zf55b08, ys93b07, and ze39f04), two from testis (qf73f12 and ot60c07), one from brain (ym44b01) and one from a pooled cDNA library (op39b10). Expression analysis of this cluster detected abundant expression in retina and cerebellum (data not shown). The other cluster, Hs.237687 (Lab ID: A166), contained only the 3'-end of a retina clone (ze36a10), but although not included in the cluster, the 5'-EST sequence was also available (ze36a10.r1). The expression profile for this cluster indicated that the sequence is expressed specifically in retina (data not shown).Attempts to both clusters by PCR amplification with primers A166F and A65R (Fig. 37 and Table 36) failed.

Bioinformatical analyses of the genomic locus containing the A166 cluster identified an amygdale mouse clone (AW457619, Fig. 37) with high homology to the human sequence. Three of its exons overlapped with gene predictions NT_004668.455 (Genscan) and ENS271417.1 (Ensembl). Since the A166 cluster sequence mapped to the intron of the murine clone and the predictions, primers A166F2 and A166R3 (Fig. 37 and Table 36) were designed to investigate whether the predicted sequences are expressed in retina. The PCR amplification confirmed that the same transcript is found in human retina. The next step was to design primers that annealed to the extremes of the gene predictions. An 1197 bp product could be amplified with primers A166F3 and A166R4 (Fig. 37 and Table 36). To further characterize the 5'-end of the gene, five different 5'-RACE experiments were done. They included use of different techniques (5'-RLM RACE and Marathon cDNA amplification) and use of various oligonucleotides and retina RNAs. Of all methods, the Clontech Marathon cDNA was the only one which produced specific product. A 580 bp product could be amplified after a primary PCR with the gene-specific primer A166F4 (Fig. 37 and Table 36) and AP1 (Table 35) followed by a nested PCR with primers A166F6 (Fig. 37 and Table 36) and AP2 (Table 35) primers. Sequencing revealed that this sequence prolonged the sequence by 172 bp (Fig. 37, L20-8). No in-frame stop codon could be found in the sequence, and although new 5'-RACE experiments with primers located in the 5'-end sequence (e.g. A166F8) were attempted, it was not possible to prolong the sequence any further. The next in-frame stop codon would be 16 bp upstream and there is only one splice acceptor site in-between. Since this splice site has splice acceptor score of 18.1 it is improbable that it will be used because 99% of all splice acceptor sites have a score of 14.0 or less (Berg and von Hippel 1998, Penotti 1991).

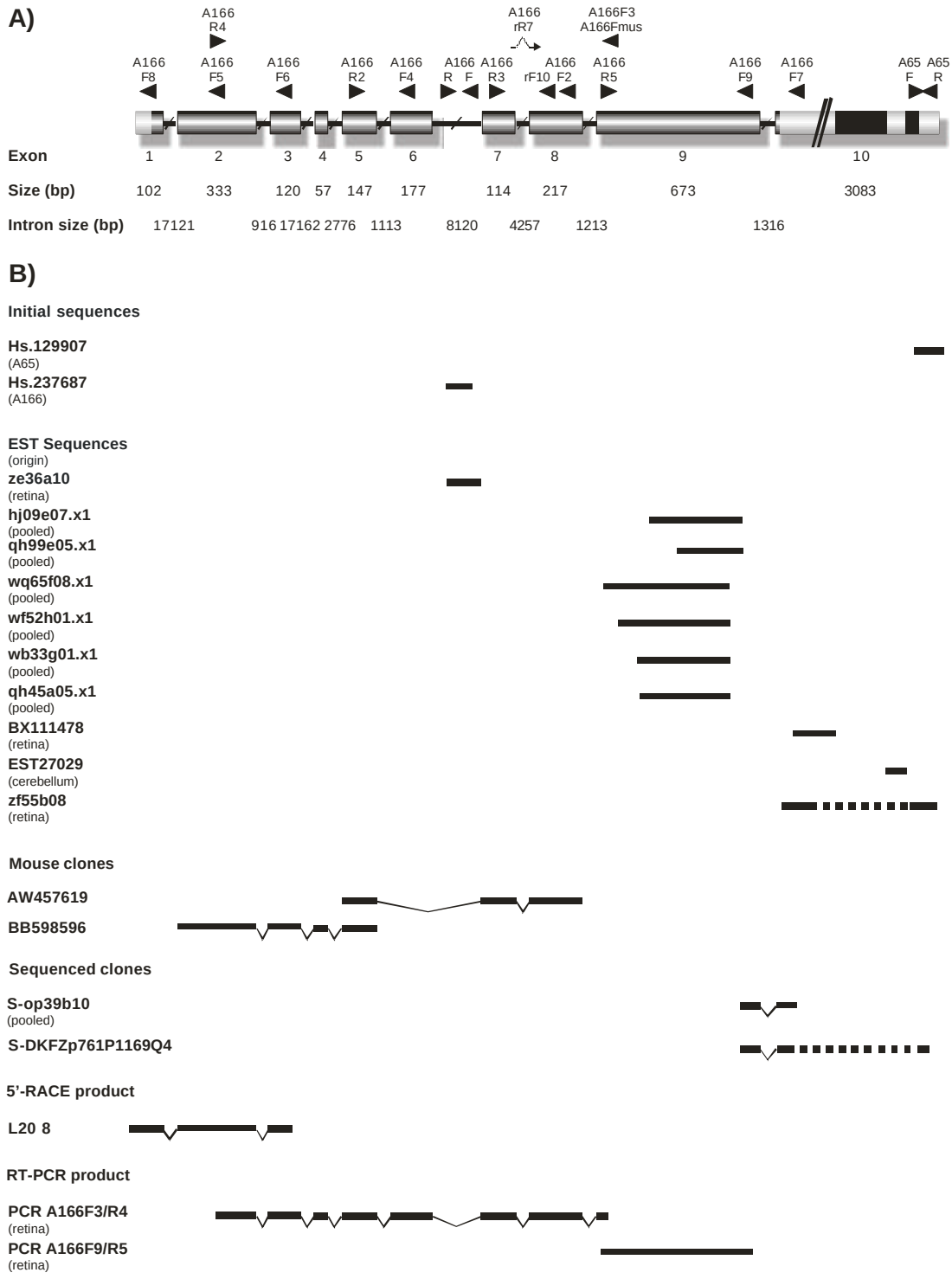


Fig. 37 Schematic representation of C1orf32

A) The 5023 bp cDNA sequence of C1orf32 is organized in 10 exons which span 59 kb of genomic sequence. The organization of the gene is shown proportional to original size for all exons except exon 10 and introns; the correct size of each is noted beneath the diagram. The open reading frame stretches from exon 1 to exon 10 and is shown as a grey box. The light grey boxes depict the 5'- and 3'-UTRs. The black shaded regions have not been sequenced from any cDNA clone. Oligonucleotide primers used in the cloning process are given above the respective exons (arrowheads).

B) The sequences of clones and PCR products which were key to the assembly of the full-length cDNA of C1orf32 are depicted. The tissue from which the mRNA sequences were obtained is listed under the name of the clone. Two clones (DKFZp761P1169Q4 and zf55b08) have not been sequenced entirely, but their insert size is known and based on this evidence it is likely that the sector which is not sequenced does not splice. This is indicated by the broken line.

In an effort to isolate clones containing the full-length gene sequence a number of library screenings were done. No specific clones could be identified by screening with a probe amplified with primers A65F and A65R (Fig. 37 and Table 36) in either the Nathans retina library or a retina library constructed in the ?TriplEx2 vector. Screening of a third retina library with probes amplified with primers A166F3/A166R2 and A166F9/A166R5 (Fig. 37 and Table 36) also failed to identify clones containing the gene sequence. By PCR amplification of 12 different cDNA library pools with A65F/A65R we were able to identify two positive clones in the brain amygdala pool and one in the hippocampus library. Although 13 different cDNA library pools were screened using two different primer combinations (A166F2/A166R3 and A166F6/A166R4) no positive clones of this region of the gene were found.

The 3'-end sequence of the gene was investigated by sequencing the 0.8 kb insert of clone op39b10 (from Hs.129907) and the amygdala clone identified from the cDNA pool screening (DKFZp761P1169Q4). Although clone DKFZp761P1169Q4 does not contain the full-length sequence, it was very informative since its 3'-end overlaps with the A65 sequences and extends the sequence 3.0 kb upstream. The 3'-UTR of the gene was assembled from these sequences, three other ESTs, namely BX111478 (retina), EST27029 (cerebellum), and zf55b08 (retina), and the genomic sequence.

Since the 5'-end of the DKFZp761P1169Q4 clone was just 503 bp downstream of the sequence that had been assembled so far, the next step was to link both sequences in order to prove that both cDNAs belong together and obtain the full-length sequence of the gene. This was achieved by amplification of retina cDNA with primers A166F9 and A166R5 (Fig. 37 and Table 36). Additional evidence that this fragment is transcribed comes from six ESTs from pooled libraries (Fig. 37).

In summary, the use of different approaches made it possible to assemble the 4954 bp C1orf32 cDNA sequence (Fig. 38). The 5'-end of the sequence was originally identified as A166 in our screening and its 3'-end as A65. From the full-length sequence, 2534 bp were sequenced in this project and can be retrieved under accession number AF503509 at the nucleotide database from NCBI. As discussed earlier, the 3'-UTR sector has not been entirely sequenced, but there is plenty of evidence to support its inclusion in the sequence. The existence of only one polyadenylation signal 50 bp upstream of the last base of the assembled sequence provides additional evidence about the assembled sequence is complete.

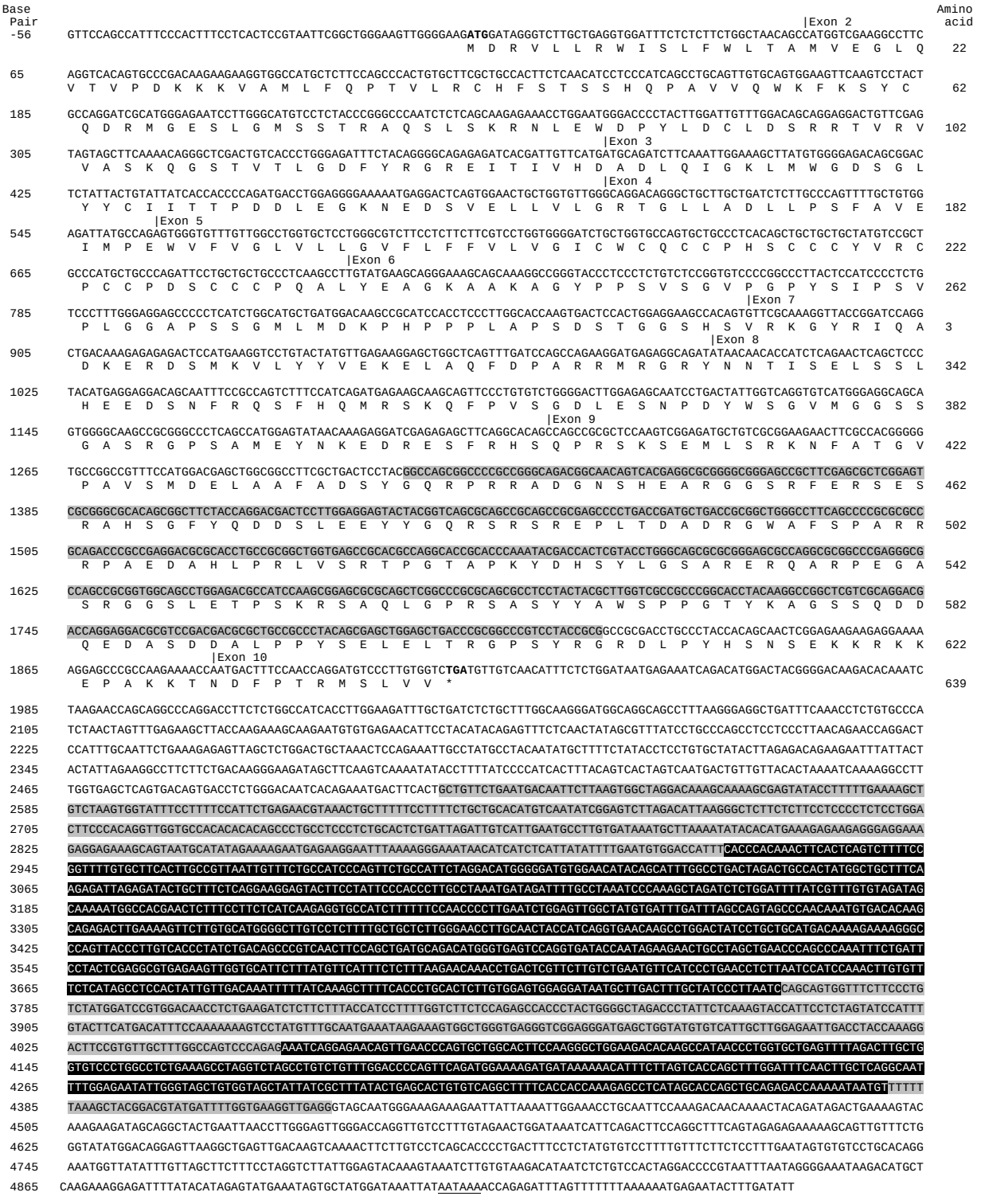


Fig. 38 cDNA and protein sequence of C1orf32

The putative cDNA sequence of C1orf32 comprising 4954 bp is shown along with the putative protein. The numbers on the left indicate the nucleotide position, the numbers on the right the aminoacid position. The bases coding for the start and stop codon are depicted with bold letters, the beginning and end of each exon is depicted by the vertical line with the corresponding exon number annotation. The cDNA sequence which has not been sequenced in this work is indicated with grey background. The light grey background delimits a region of exon 9 which is supported by numerous ESTs from pooled libraries. The grey shaded sequence in exon 10 is supported by three ESTs, namely BX111478 (retina), AA324135 (cerebellum), and AA058586 (retina). The black background identifies two regions for which there is only indirect EST evidence.

6.4.2. Genomic structure of C1orf32

Alignment of the full-length cDNA to genomic clone AL009182 from chromosome 1q24.1 revealed that the gene is composed of 10 exons covering approximately 60 kb of genomic DNA. The exon sizes range from 57 to more than 3000 bp (Table 28).

Table 28 Exon / intron structure of human C1orf32

Exon		3'-Acceptor Splice Site[a]		5'-Donor Splice Site[a]		Intron	
No.	Size (bp)	Sequence	Score[b]	Sequence	Score[b]	No.	Size (bp)
1	102			CAGgtaaga	1.06	1	17,121
2	333	ttattctgttttgcagC	4.86	ATGgtaata	5.65	2	916
3	120	ctttttgtgataacagA	7.26	TGGgtaagt	2.64	3	17,162
4	57	cttccacttcttctagG	4.38	CAGgtattc	6.14	4	2776
5	147	ttgtgcatccttccagA	5.71	CCTgtgagt	3.76	5	1113
6	177	tgatctccctccacagT	5.03	GTGgtaaat	4.81	6	8120
7	114	gtgcttttctcaacagT	6.02	GATgtgagc	4.11	7	4257
8	217	gattttgtccctgcagA	5.01	CAGgtgatg	3.77	8	1213
9	673	ccttctctgcccgcagC	3.69	ACCgtgaga	5.80	9	1316
10	3083	ttctgctctttcctagA	4.39				

[a] Exonic and intronic sequences in upper and lower case letters, respectively.
[b] Score of donor/acceptor splice site. According to published data (Berg and von Hippel 1998 and Penotti 1991) 99% of sites have a score of 0-11 (donor) or 0-20 (acceptor). Scores were calculated using the spreadsheet created by Christian Sauer, 2001.

6.4.3. *In-silico* assembly of the mouse C1orf32 orthologue

The cloning process of the human C1orf32 gene was greatly aided by the existence of the AW457619 murine clone, which localizes to murine chromosome 1. In order to evaluate the existence of a murine homologue of C1orf32, the human and murine genomic contigs containing C1orf32 and AW457619, were compared. This genomic interspecies comparison and clones AW457619, BM935257, BB598596, BY727699, and BY727657 were used to assemble the coding sequence of the murine C1orf32 gene. With the addition of the 3'-UTR, which was assembled from more than 40 ESTs extending more than 6kb downstream of the assembled sequence the C1orf32 cDNA consists of a 8254 bp transcript.

The murine gene is also organized in 10 exons (Table 29). Exons 2, 3, 4, 5, 6, and 7 have the same length in both species. The 5'-end of the mouse gene, which is based on two clones (BY727699 and BY727657) from a full-length enriched library from neonate medulla oblongata which were made public in December 2002, is 266 bp longer than the human. Based on EST evidence, the murine 3'-UTR is 3 kb longer than the human. The murine 3'-UTR contains eight potential polyadenylation signals, with the last located just 18 bp from the end of the putative sequence.

Translation of the murine C1orf32 gene results in a putative protein of 661 aa, which is 22 aa longer than the human counterpart (Fig. 41). The same domains predicted for the human protein are also found in the mouse protein.

Table 29 Hypothetical exon / intron structure of murine C1orf32

Exon		3'-Acceptor Splice Site[a]	5'-Donor Splice Site[a]	Intron	
No.	Size (bp)	Sequence	Sequence	No.	Size (bp)
1	328		CAGgtaaga	1	14,838
2	333	tccttctgttttacagC	ACGgtaatg	2	961
3	120	ctttctgtgataatagA	TGGgtgagt	3	20,798
4	57	cttctactccttctagG	CAGgtattc	4	3118
5	147	ttatgcatccttccagA	CCTgtgagt	5	1014
6	177	tcatctccctccacagT	GTGgtaatt	6	7573
7	114	gtgctttctccaacagT	GATgtgagc	7	4104
8	214	gcttttgttcctgcagA	CAGgtgacg	8	913
9	698	gaccatctgcccgcagC	CCCgtgagg	9	1216
10	6066	ttttgcttttccctagG			

[a] Exonic and intronic sequences in upper and lower case letters, respectively.

6.4.4. Expression analysis of C1orf32

The expression of C1orf32 was analyzed several times in the course of the investigation. The original amplifications done with primers A166F/A166R revealed a retina-specificity; amplification with A65F/A65R showed expression in retina, brain, and kidney. Abundant expression of the transcript in retina, with some expression also observed in thymus, brain, heart, spinal cord, testis, colon, and placenta was confirmed by amplifications with primer combinations A166F2/A166R3, A166F3/A166R4 and A166F3/A166R3 (Table 36, Appendix).

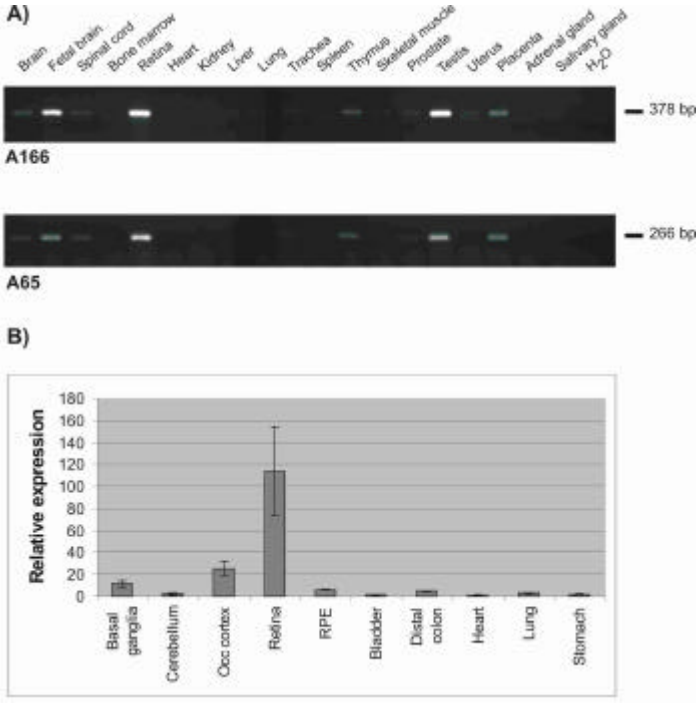


Fig. 39 Expression of C1orf32 and its isoforms

A) Expression profiling of the 5'- (A166) and 3'-ends (A65) of C1orf32 reveals that the gene is predominantly expressed in retina with less expression in fetal brain, testis, brain, placenta, spinal cord, thymus, prostate, and uterus. There is a second product in uterus which was not futher analyzed and may represent a second splice variant.
B) The expression of C1orf32 was confirmed by qRT-PCR in a panel of 10 tissues. Amplification of an 85 bp product with primers A166rF10 and A166rR7 confirms that the C1orf32 transcript is expressed preferentially in retina. The expressions in bladder, distal colon, lung and stomach are slightly overestimated due to detection of another product by the SYBR-Green dye.

The expression profile determined with the A166 and A65 primers provided additional evidence that both fragments derive from the same gene as they show overlapping expression. To prove this, the

same 19-tissue panel was amplified with primers A166F4/A166R4 and A65F/A65R (Fig. 39A and Table 36). As can be seen, both primer sets amplify a product from the same tissues, with a clear correlation in the intensity of expression for both PCR reactions.

The expression was quantified by qRT-PCR using primers A166rF10 and A166rR7 (Fig. 36 and Table 37). From the tested tissues, retina reveals the highest expression, followed by occipital cortex, and basal ganglia (Fig. 39B). Even though very low expression can be seen in the non-neuronal tissues, the level is even lower than that represented in the figure. This is due to the fact that there are some unspecific products amplified in this tissues and the SYBR-Green detection also measures these secondary products.

To asses the length of the transcript and to investigate the expression of the murine gene, a fragment of the 5'-end of the murine gene was amplified with primers A166R3 and A166Fmus (Table 36). Hybridization of the radio-labeled product to a Northern blot containing mouse brain, eye, heart and lung resulted in specific hybridization to a transcript over 6.5 kb in size (Fig. 40). The signal was stronger in brain and eye, but it was also visible in heart and lung.

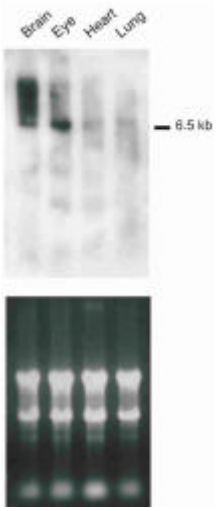


Fig. 40 Northern blot analysis of the murine C1orf32 transcript
Total RNA isolated from mouse brain, whole eye, heart, and lung RNA was separated electrophoretically in a formaldehyde-containing agarose gel (lower picture) and blotted. Hybridization of the filter with a PCR product amplified from murine eye with primers A166R3 and A166Fmus revealed strong expression in eye and brain and to a lesser degree in heart and lung (upper picture). It must be kept in mind that the expression in retina is probably higher than observed, since the RNA of the whole eye was hybridized.

6.4.5. *In-silico* analysis of the putative C1orf32 protein

Exons 1 thru 10 of human C1orf32 encode a hypothetical protein of 639 aa with a molecular weight of 71.2 kDa and an isoelectric point of 8.43. The mouse orthologue translated from the *in-silico* assembled transcript is 651 aa long and has an 87% similarity to the putative human protein (Fig. 41). A BLAST search with the human protein also identified protein sequence XP_222845, which has been predicted from the rat genome by automated computational analysis (GenomeScan[63]) and contains partially the same ORF as C1orf32. The rat prediction contains an extra exon between exons 6 and 7 and 7 and 8 of human C1orf32 plus additional exons in the 3'-end. Whereas the rat and mouse orthologues have a homology of 92% in the compared region, the human/rat similarity is 85%.

[63] http://genes.mit.edu/genomescan.html

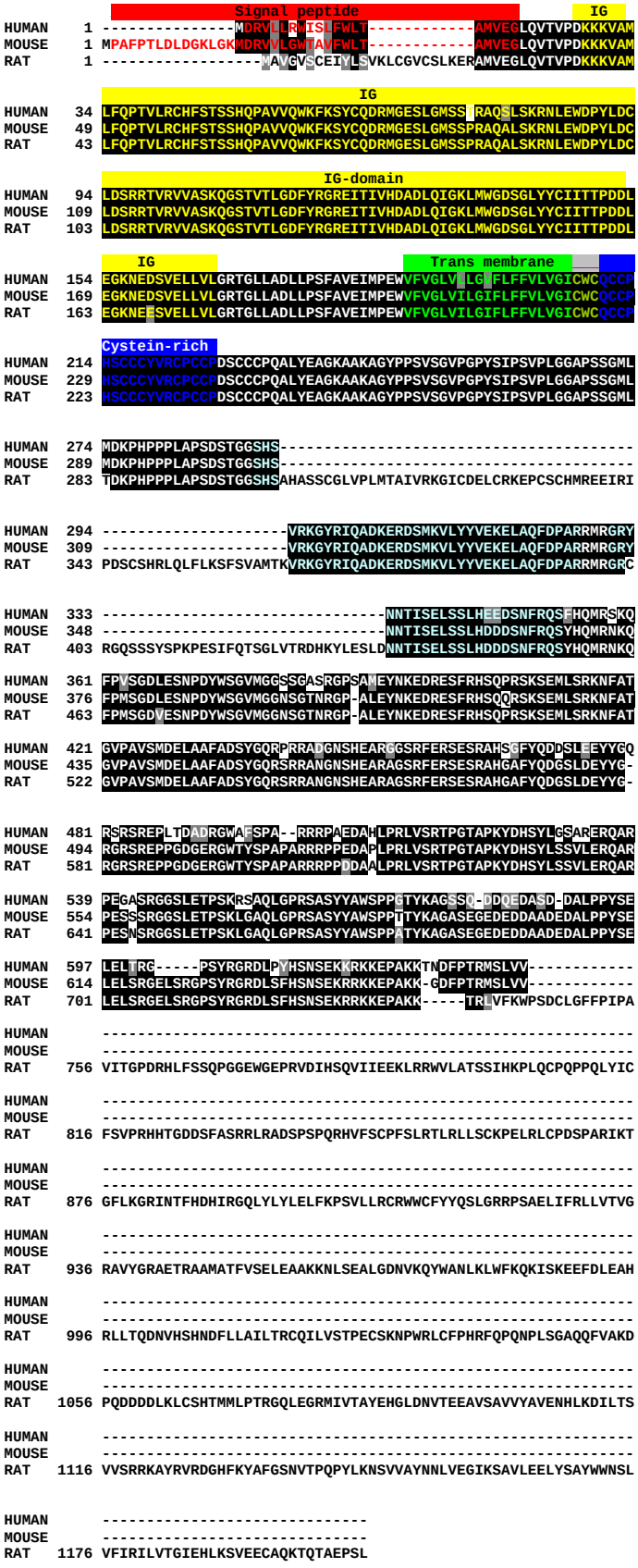


Fig. 41 Sequence alignment of the rat, murine, and human C1orf32 hypothetical proteins

The human and murine C1orf32 putative proteins are based on mRNA and EST evidence. The rat sequence is based on prediction from genomic sequence (XP_222845). As can be observed, the majority of amino acids is highly conserved in the three species (identical amino acid residues shaded in black, grey shade indicates those with similar chemical properties and/or structure). The immunoglobulin (IG), transmembrane (TM) domain and cystein-rich region are also found in all three species. Three bases of the transmembrane domain overlap with the cystein-rich region. A signal peptide of 20 aa and 35 aa in the human and mouse protein, respectively is not found in the rat sequence. The light blue amino acids show regions which might constitute novel motifs of the protein family. The long stretches of extra sequences observed in the rat protein are contained in additional exons found in the prediction.

The common characteristic of the proteins from the three species is the presence of an immunoglobulin (IG) and a transmembrane (TM) domain (Fig. 41). Another particularity is the

presence of a cystein-rich region found in all proteins. Post-translational modification of the proteins includes the cleavage and release of a signal peptide of 20 (human) and 35 aa (mouse), respectively. The signal peptides differ, because the first peptides of the protein are encoded in exon 1 which is different in all three species. No signal peptide is predicted to be cleaved from the rat protein.

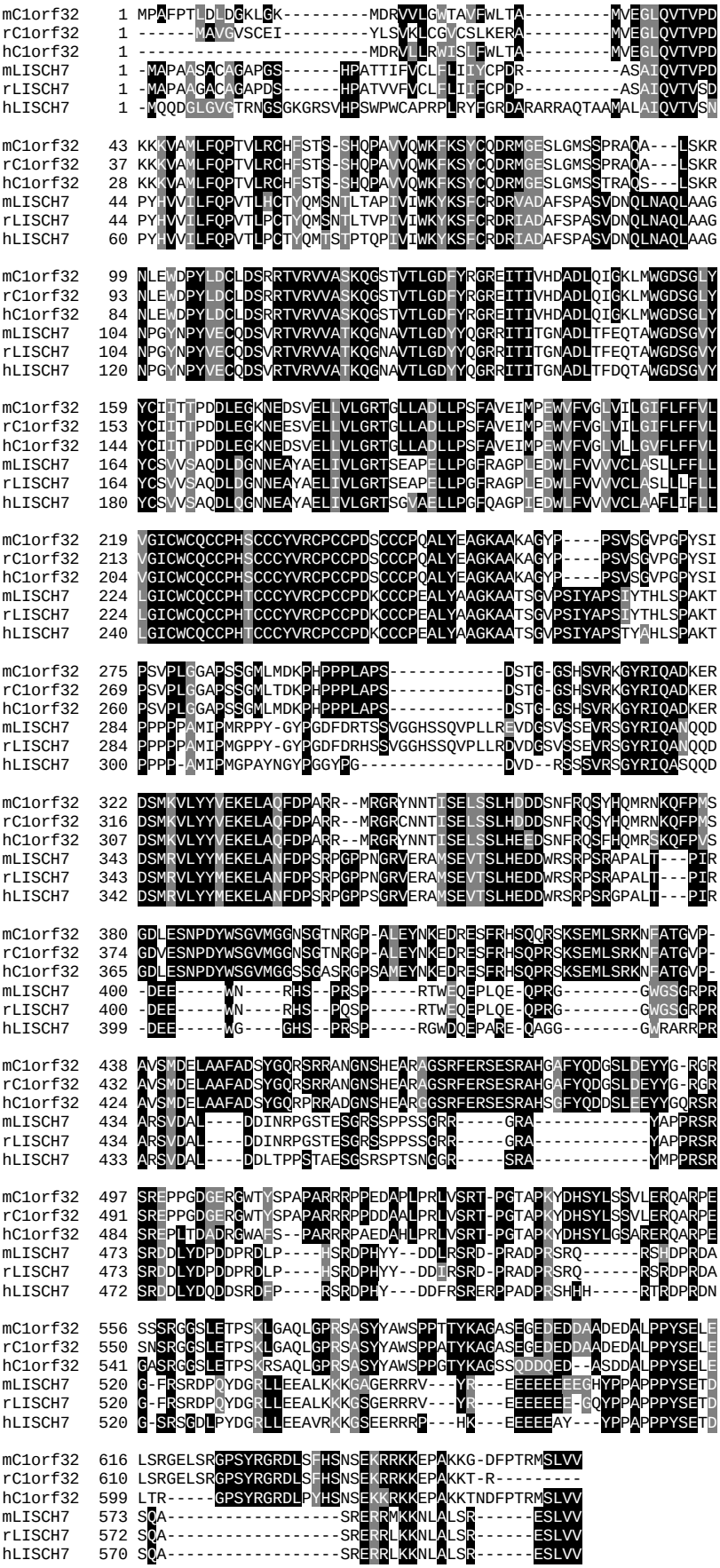


Fig. 42 Sequence alignment of the human, murine, and rat LISCH7 and C1orf32 proteins

The mouse (m), rat (r), and human (h) LISCH7 and C1orf32 proteins were aligned. As can be observed, both genes share segments of high homology. Some of these have already been identified as domains (cystein-rich, immunoglobulin, transmembrane). Others are not found in any domain or motif database.

The homology searches also identified another gene which might belong to the same family as C1orf32. The gene, called liver-specific bHLH-Zip transcription factor (LISCH7) has been reported in all three species (GenBank acc. no. NP_057009, NP_059101, and NP_116005) and is highly

conserved. The transcripts of both genes from all three species were compared using the BLOSUM62 matrix[64] (Fig. 42). The comparison shows that there may be additional novel motifs which are present in both genes in all species but have not been described as such in the public databases until now. To generate the three proposed motifs, segments of high homology were analyzed with the PRATT tool[65], which generated the following patterns S-x(0,1)-V-R-x-G-Y-R-I-Q-A-[DNS]-x-[EQ]-[DR]-D-S-M-[KR]-V-L-Y-Y-[MV]-E-K-E-L-A-[NQ]-F-D-P-[AS]-R, K-[KR]-x-K-K-[EN]-[LP]-A, and G-R-x-[EN]-[NR]-[AT]-[IM]-S-E-x(1,2)-S-(0,1)-L-H-[DE]-[DE]-D-x-[NR]-x-R-[PQ]-S. Additionally, at the 3'-end of all sequences an S-L-V-V block is observed and thus probably has a functional significance. The Pfam database[66] contains a motif for the LISCH7 family, which is found under accession number PF05624.

In order to establish the relationship between both genes and the different species, a phylogenetic tree was created after alignment of the six sequences using the CLUSTALW algorithm[67] (Fig. 43).

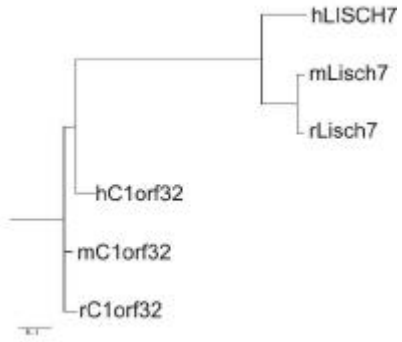


Fig. 43 Phylogenetic tree of C1ORF32 and LISCH7
The aligned sequences include NP_057009 (human), NP_059101 (mouse), and NP_116005 (rat) for LISCH7. The human C1orf32 sequence can be found under acc. nos. AF503509. The rat sequence is based on sequence XP_222845, but the extra exons have been deleted.

6.5. Cloning and characterization of C14orf29 (B015)

6.5.1. Assembly of the cDNA sequence of C14orf29

The cloning of this gene was initiated when Hs.194617 and its TIGR counterpart THC826685 were shown to be preferentially expressed in retina during the second phase of the UniGene project. The cluster contained 13 clones, most of them sequenced from pooled germ cells and just one which originated from retina (yt04g11). Therefore, the B15F and B15R primers (Fig. 44 and Table 36) were designed to anneal to the exons contained within the retina clone in order to evaluate the expression of this gene in retina and other tissues. Since the gene was highly expressed in retina, efforts were made to assemble the cDNA sequence of the transcript. Translation of the TIGR assembly revealed that the compilation probably contained the complete 3'-end of the gene, since the putative ORF of 252 aa was followed by a stop codon and the nucleotide assembly contained a polyadenylation signal just 26 bp upstream of the mRNA 3'-end. Analysis of the 5'-end of the assembly revealed that it was probably incomplete, as it did not contain an in-frame stop codon. The existence of a prediction by Genscan (NT_025892.433) also supported the hypothesis of additional 5'-exons. In order to analyze if

[64] http://www.ebi.ac.uk/clustalw/index.html
[65] http://www.ebi.ac.uk/pratt/
[66] Pfam: http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF05624
[67] http://www.ebi.ac.uk/clustalw/index.html

the transcript present in retina contained the same exons as the pooled cell germ transcripts, primer B15R2 (Table 36) was designed to anneal at the end of the known 3'-UTR. A PCR using primers B15F/B15R2 (Table 36) resulted in an approximately 1.5 kb product. An aliquot of the PCR was cloned and 16 clones were picked and amplified with vector primers (M13F/M13R). Sequencing of four selected clones (T32-4, T32-5, T32-7, and T32-14) revealed that retina transcripts do not contain one of the exons found in the THC826685 sequence (Fig. 44). The present assembly, THC265755 does not include the clone which encoded this exon. The smaller insert found in clone T32-7 results from the use of an alternative splice donor which removes 112 bp contained in the other transcripts (Fig. 44).

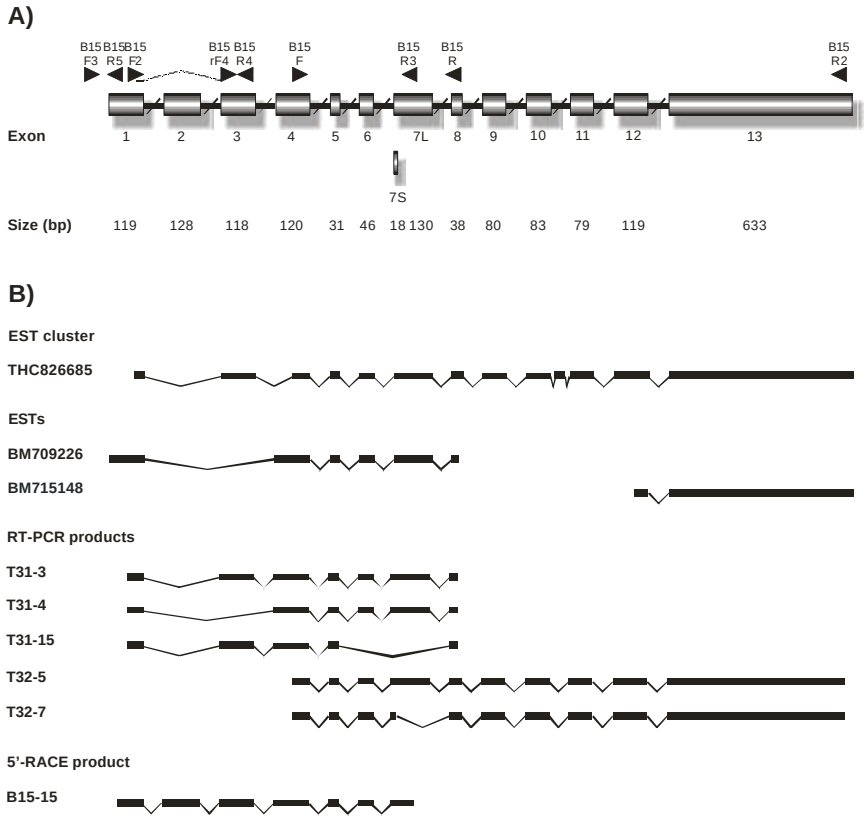


Fig. 44 Schematic representation of C14orf29
A) The 1742 bp of the C14orf29 gene are organized in 13 exons which are represented by grey boxes proportional to original size (noted below each). Two alternative splice forms are known for exon 7 (Exon 7L and 7S) and four of the exons (2, 3, 6, and 7) are spliced out in some transcripts.
B) Cloning of the gene was achieved by a number of PCR and 5'-RACE experiments were carried out using the primers depicted above the exons. The resulting products were cloned and in the figure a selection of sequences used for the assembly of the C14orf29 full-length sequence are depicted. Clones BM709226 and BM715148 have been sequenced from an optic nerve and retina library, respectively. They were made public after the full-length sequence had been assembled and prolonged by some bases the 5'- and 3'-ends.

The effort to identify novel 5'-exons by PCR amplification with oligonucleotides B15F2/B15R was successful. Using retina cDNA as template four different products were amplified (Fig. 44). As can be seen in Fig. 45, the amplification of each isoform depends on the $MgCl_2$ concentration. The PCR products obtained using 1.0 mM $MgCl_2$ were cloned and five clones (T31-3, T31-4, T31-11, T31-14, and T31-15) were sequenced. These clones represented three different splice variants corresponding to the 318, 391, and 494 bp bands observed in the gel (Fig. 45). Based on the sizes of the inserts, none of the clones analyzed contained the bigger product, but its sequence could be determined after identification of an additional exon in a 5'-RACE product. Thus, by PCR amplification it was possible to extend the 5'-end of the gene by 388 bp and to identify four novel splice variants (Fig. 44).

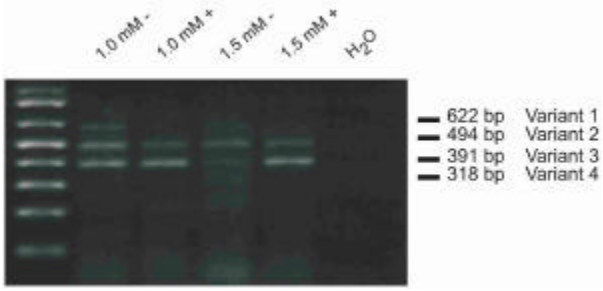


Fig. 45 RT-PCR amplification with primers B15F2/B15R
The analysis of the different isoforms was carried out by amplification of retina cDNA using different $MgCl_2$ concentrations (1.0 or 1.5 mM) and presence (+) or absence (-) of formamide. Even though the 318 bp is not visible with the EtBr staining, after cloning of the product of the reaction done with 1.0 mM $MgCl_2$ in the absence of formamide and sequencing of clones, many clones containing variant 4 were found whereas no clones containing variant 1 were identified.

Since the Genscan prediction NT_025892.433 extended somewhat upstream of where the B15F2 primer was located and that the products amplified from the 5'-end did not contain an in-frame stop codon, 5' RACE was done. Marathon retina cDNA was amplified in a first PCR with primers B15R and AP1 (Table 35). A 1:100 dilution of this product was then amplified with B15R3 and AP2 (Table 35) at various $MgCl_2$ concentrations with and without formamide. The product obtained in the secondary PCR was cloned and twenty-five colonies were picked. Sequencing of clones B15-14, B15-15, B15-20, B15-21, B15-22, and B15-13 revealed that four of the products did not extend to the known 5'-sequence. Only B15-15 (Fig. 44) extended up to the first exon but ended downstream of primer B15F2. Nevertheless, the clone sequence B15-15 was the first evidence of the existence of 128 bp exon located between exons 1 and 3. This is also the additional sequence found in the largest product obtained by amplification with B15F2/ B15R (Fig. 44 and Table 36).

Since the 5'-RACE amplifications did not prolong the 5'-end sequence primers B15F3 and B15R5 (Fig. 44 and Table 36) were designed to anneal upstream of the hypothetical in-frame stop codon and at the 3'-end of the same exon, respectively (Fig. 44). The sequence encompassed by the primers has a high GC content and it was not possible to optimize the conditions in order to amplify specific products either with primers B15F3/B15R4 or B15F3/B15R5 (Fig. 44). Even though it has not been possible to generate a clone which contains the 5'-UTR region of the gene, there is independent evidence to suggest that the assembled transcript contains the entire coding sequence of C14orf29. This evidence comes from clones BM709226 and BM709207 derived from optic nerve and BX096176 and H86299 derived from retina. All except BM709226 extend a few bases upstream of the sequence which could already be amplified in our project. Clone BM709226 contains 84 additional bases of the 5'-end, but the in-frame stop codon is still 99 bp upstream of this end. The three gene predictions (SGP, Geneid, Genscan) identify the start of the transcript at this position. Similarly, the FirstEF prediction tool[68] does not predict any other upstream exon. It is therefore reasonable to assume that the full-length coding region and the partial sequence of the 5'-UTR of C14orf29 have been cloned.

Based on the sequenced PCR products and reported ESTs, four different variants of the gene have been assembled (identified as C14orf29_v1, _v2, _v3, and _v4). Their GenBank accession numbers are AY311396, AY311397, AY311398, and AY311399. The longest transcript, C14orf29_v1, is 1708 bp long. At least one more variant, which contains exon 7S, is expressed. In spite of this fact, no variant has been assembled since this alternative splice was found only in a single product (T32-7)

[68] http://rulai.cshl.org/tools/FirstEF/

which did not include the 5'-end of the gene. Additionally, none of the products amplified from the 5'-half of the gene (with primers B15F2/ B15R) contained exon 7S. Exon and total length of variants 1 to 4 are detailed in Table 30.

6.5.2. Genomic structure of C14orf29

The full-length sequence of C14orf29 spans 32 kb and is composed of 13 exons (Table 30) mapping to human chromosome 14q22.1. The 3'-UTR of the gene is located just 250 bp upstream of the 3'-UTR of the phosphorylase, glycogen, liver (PYGL) gene, which is transcribed from the other strand.

The exons range in size from 18 to 632 bp and all splice sites comply with the donor-acceptor conserved sequences. Four of the exons (nos. 2, 3, 6, and 7) are alternatively spliced in some transcripts. Exon 7 also contains an alternative donor site. The two ensuing variants of exon 7 are identified as 7S (shorter exon) and 7L (longer exon).

Table 30 Exon / intron structure of C14orf29

Exon		3'-Acceptor Splice Site[a]		5'-Donor Splice Site[a]		Intron	
No.	**Size (bp)**	**Sequence**	**Score[b]**	**Sequence**	**Score[b]**	**No.**	**Size (bp)**
1	119			GCGgtgagt	2.38	1	5673
2	128	tcccaataccttgcagA	7.67	TTTgtaagt	5.04	2	665
3	103	gatttttactttccagT	5.72	CTGgtgagt	1.56	3	1604
4	120	ctcccttgctcttcagG	2.37	CAGgtcagt	3.34	4	1024
5	31	taaacttccttcacagG	5.00	AAGgtatgt	2.55	5	3987
6	46	ctcctttgttacctagG	5.06	GAGgtatgt	3.19	6	106
7L	130	cttttcttgctgccagG	3.67	AGGgtaagt	1.55	7L? 8	751
7S	18	cttttcttgctgccagG	3.67	CAGgtaagc	0.93	7S? 8	863
8	38	tttcattgctttacagA	4.69	AAGgtaata	4.17	8	2139
9	80	tttctttcatttcaagG	5.14	AAGgtgaga	1.36	9	12,925
10	83	tctttttaaaaatagA	9.24	AAAgtaagt	2.31	10	1477
11	79	ttctacttgtccctagT	6.66	AAGgtaaac	3.61	11	606
12	119	tttctgtcgcttgcagC	4.91	GAGgtaaga	1.92	12	146
13	632	ggtccttttctacagA	3.52			13	5673

[a] Exonic and intronic sequences in upper and lower case letters, respectively.
[b] Score of donor/acceptor splice site. According to published data (Berg and von Hippel 1998 and Penotti 1991) 99% of sites have a score of 0-11 (donor) or 0-20 (acceptor). Scores were calculated using the spreadsheet created by Christian Sauer, 2001.

6.5.3. Expression analysis of C14orf29

To investigate the expression of the C14orf29 variants, PCR with primers B15F2 and B15R (Fig. 44 and Table 36) was done in a 20-tissue cDNA panel (Fig. 46). The retina is the only tissue where all isoforms are expressed. In uterus very low expression of variant 4 is observed.

The expression profile was confirmed by qRT-PCR using primers B15rF4 and B15R4 (Fig. 44 and Table 37) in a panel of ten cDNAs from various tissues. Since 11 bp of primer B15rF4 anneal to exon 1 and 6 bp to exon 3, only those isoforms which do not include exon 2 were amplified. The 114 bp

product from isoforms 3, 4, and 7 was found to be most abundantly expressed in occipital cortex, followed by retina, and basal ganglia (Fig. 46). The expression in RPE and cerebellum was significantly lower and no expression was observed in the non-neuronal tissues.

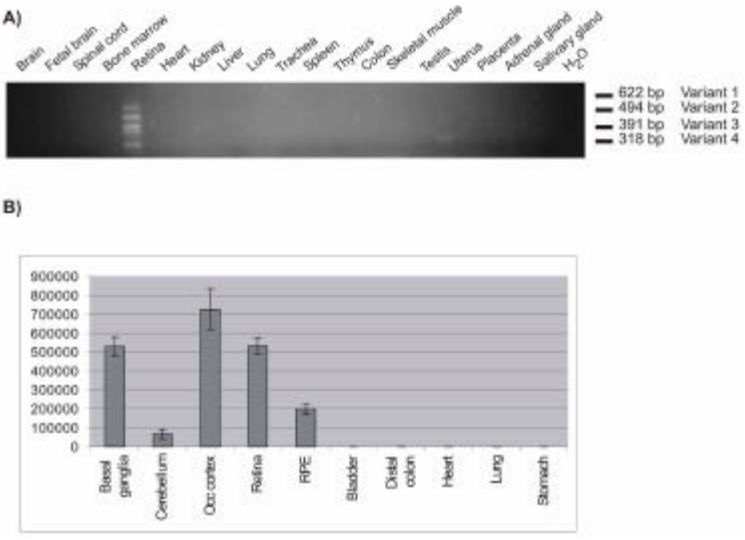


Fig. 46 Expression of C14orf29 and its isoforms

A) The expression profile of C14orf29 and its isoforms was determined by amplification with primers B15F2/B15R in a panel of 20 tissues. Four different products were amplified from retina; only variant 4 was also weakly amplified in uterus.

B) The expression was confirmed with qRT-PCR using primers B15rF4/B15R4 which amplify only the variants without exon 2. As can be seen in the diagram, the obtained product is expressed exclusively in neuronal tissues. It must be noted that the expression in basal ganglia, cerebellum, and occipital cortex is overestimated because of a second product amplified in these tissues which is also measured by the SYBR-Green staining.

6.5.4. *In-silico* analysis of the putative C14orf29 proteins

Translation of exons 1 to 13 of C14orf29_v1 results in a 362 aa protein with a predicted molecular weight of 40.8 kDa and an isoelectric point of 8.57. The other variants encode shorter versions of this ORF. Predicted posttranslational modifications of the protein include O-glycosylation of S11, phosphorilation of serines 11, 40, 59, 155, 182 and threonines 183 and 209, and possibly glycosylation by O-linked-N-acetylglucosamine of six different amino acids (S11 and S358 have the greatest probability). No protein cleavage, N-glycosylation, or transmembrane domains were predicted. The protein contains an abhydrolase domain stretching 60 amino acids. This alpha/beta hydrolase fold is common to a number of hydrolytic enzymes of widely differing phylogenetic origin and catalytic function (Ollis et al. 1992). The core of each enzyme is an alpha/beta-sheet (rather than a barrel), containing eight strands connected by helices.

To identify proteins which might be related to C14orf29, searches using the BLAST tool[69] were carried out. This identified a number of proteins similar to C14orf29, but only the murine XP_286429 sequence is similar to the N-terminal region. All other similarities start after amino acid 70. Most of the sequences derive from a transcript encoded on chromosome 20p11.21, whose representative sequence is NP_056415. In a segment of 282 aa there is 50% identity and 68% similarity between C12of29 and NP_056415. An interesting finding is that the 3'-end of NP_056415 is just 1.5 kb downstream of the 3'-end of the phosphorylase, glycogen, brain (PYGB) gene. This gene is part of the glycogen phosphorylases, a family of three isozymes: muscle, liver, and brain, which are expressed selectively and to varying extents in a wide variety of cell types. The enzyme encoded by this gene is the rate-limiting enzyme catalyzing glycogen degradation. C14orf29 localizes close to PYGL (section

[69] http://www.ncbi.nlm.nih.gov/BLAST/

6.5.2), which is 81% identical to PYGB. Thus, it seems that the two gene families are similarly organized on the respective chromosomes.

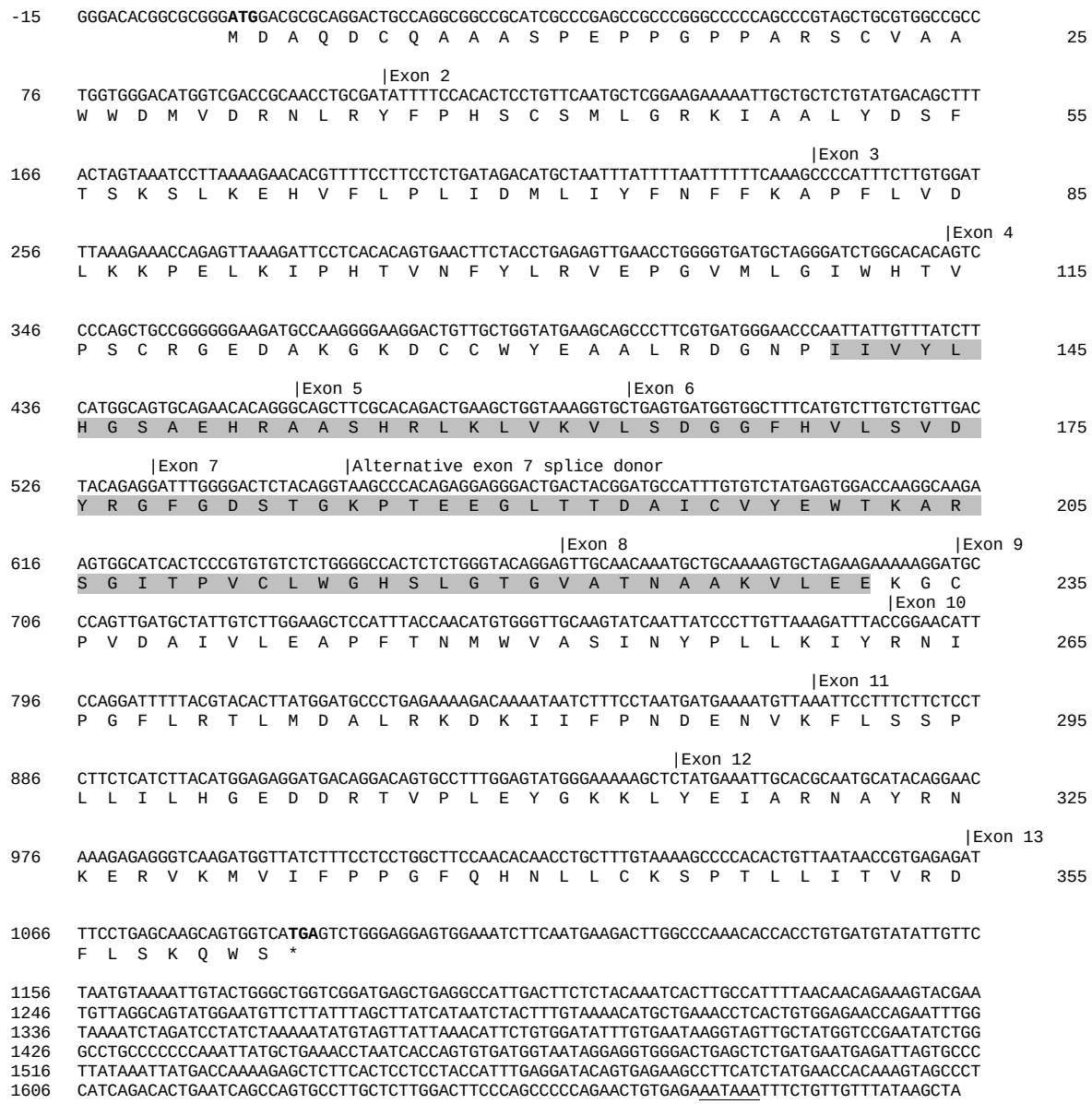


Fig. 47 cDNA and protein sequence of C14orf29_v1
The longest cDNA sequence of C14orf29 is composed of 13 exons. The start and stop codons of the C14orf29_v1 putative protein are located in exon 1 and 13, respectively (indicated by bold letters). A polyadenylation signal (underlined) is found 26 bp upstream of the 3'-end of the transcript. The encoded protein contains an esterase domain (indicated by grey-shaded amino acids).

As mentioned above, murine RefSeq XP_286429 was the only protein that had a 68% similarity to the N-terminal region of the protein. When this sequence was aligned to the mouse genomic assembly, it was observed that the last exon of the sequence overlapped with a number of gene predictions and that there is an mRNA clone (BY721299) which contains the same exons as the RefSeq sequence. Thus, the orthologous murine gene was assembled from sequence XP_286429 and ENSEMBL gene prediction ENSMUST00000061935.1[70]. Since the 5'-end of the gene prediction contains sequence

[70] http://www.ensembl.org/Mus_musculus/geneview?transcript=ENSMUST00000061935

which is absent in the mRNA clone; the first bases of the prediction were altered accordingly. The resulting 351 aa putative protein has a 73% similarity to the putative human C14orf29 protein (Fig. 48). Further evidence that human C14orf29 and the mouse assembly are orthologous comes from the fact that the PYGL gene is located just 1.5 kb downstream of the murine sequence. The BLAST search also identified rat entry XP_234268, a model reference sequence predicted from NCBI contig NW_043951 by GenomeScan[71]. Query of the rat genomic sequence using the USCS Genome Browser[72] identified prediction chr6.88.015a which is longer than rat XP_234268 and was therefore used for the comparisons with the human gene. This sequence is highly homologous to the human protein (Fig. 48), but has an alternate N-terminus and exons 5, 6, and 7 are missing. From the 114 aa in this sequence, 93% have similar chemical properties to C14orf29. This hypothetical gene is located in rat chromosome 6 and its 3'-end is also very close to the rat PYGL gene.

[71] http://genes.mit.edu/genomescan.html
[72] http://genome.ucsc.edu/cgi-bin/hgBlat

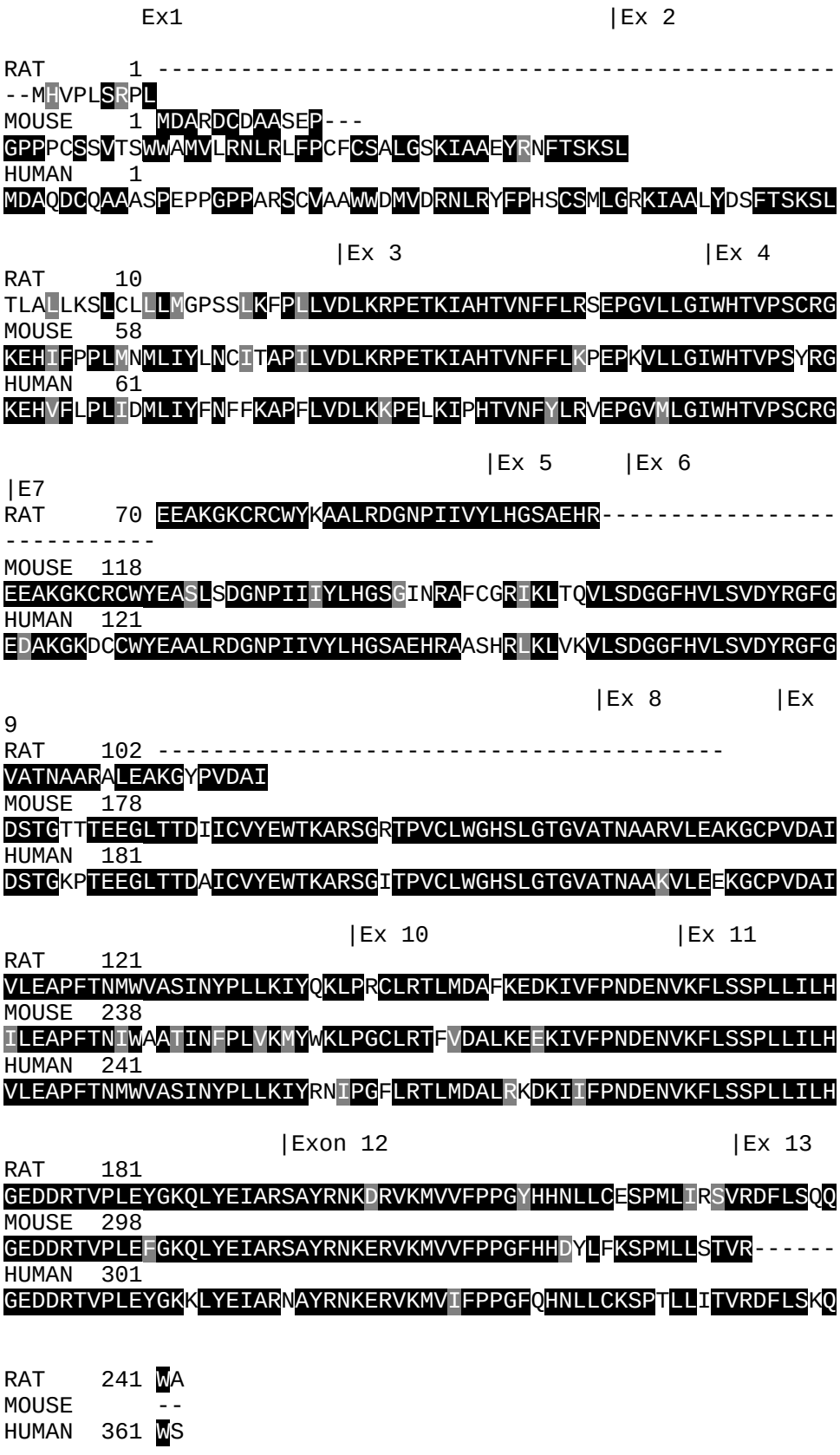


Fig. 48 Comparison of the rat, murine, and human C14orf29 proteins

The aligned human sequence was translated from C14orf29_v1. The mouse sequence was assembled from SGP prediction chr12_891.1 and XP_286429. The first peptides of the prediction have been left out in order to assemble a full-length sequence. The rat sequence corresponds to GeneScan prediction chr6.88.015.a.

6.6. Cloning and characterization of C4orf11 (L39)

6.6.1. Assembly of the cDNA sequence of C4orf11

Clone 3A12F was identified from the retSSH library. Its 473 bp sequence spliced twice and comparison to the public cDNA sequences determined that it was very similar to clone ys80g07 which also derives from retina (Fig. 49). The only difference between the two sequences was an insertion of 127 bp in clone ys80g07 which corresponds to a complete exon. Other evidence of a gene at this locus of chromosome 4 came from GenScan prediction NT_016354.148 which assumes 4 exons, but only overlapped with the library clone in the second exon. To clone the full-length sequence a mix of adult human retina libraries C1F1, DKFZ3, and DKFZ4 was hybridized with a probe obtained by amplification with primers L39F/L39R (Fig. 49 and Table 36). From the approximately 1 x 10^6 pfu screened in duplicate, 11 clones (identified as 39-1-1, 39-1-2, 39-1-3, 39-3-1, 39-3-2, 39-4-1, 39-5-1, 39-6-1, 39-6-2, 39-6-3, and 39-6-4) were further analyzed. These ?TriplEx2 clones were converted to pTriplEx2 plasmids which had inserts between 0.8 and 2 kb. The isolated plasmid DNA of each clone was subsequently sequenced with primers ?TriplEx 5', ?TriplEx 3', L39F, L39R, L39F3, L39F4, and L39R3 (Fig. 49 and Table 36). Analysis of the sequences revealed that most of the clones contained alternatively spliced transcripts, and in many cases the introns were retained (Fig. 49). PCR amplification of retina cDNA with primers L39F2 and L39R (Fig. 49 and Table 36) amplified two products of 596 (L39F2/Ra) and 658 bp (L39F2/Rb). These products contained the same sequence already found in the library clones. The amplification with L39F and L39R2 (Fig. 49 and Table 36) amplified only one product of 430 bp even though library clone 39-5-1-1 contains 487 additional base pairs between these two primers. The extra sequence found in clones 39-1-3 and 39-3-1-1 could be verified by amplifications using primers L39F3/L39R3 (Table 36) which amplified a 1152 bp product. Thus, it seems that the intronic sequence found in some library clones is due to incomplete splicing.

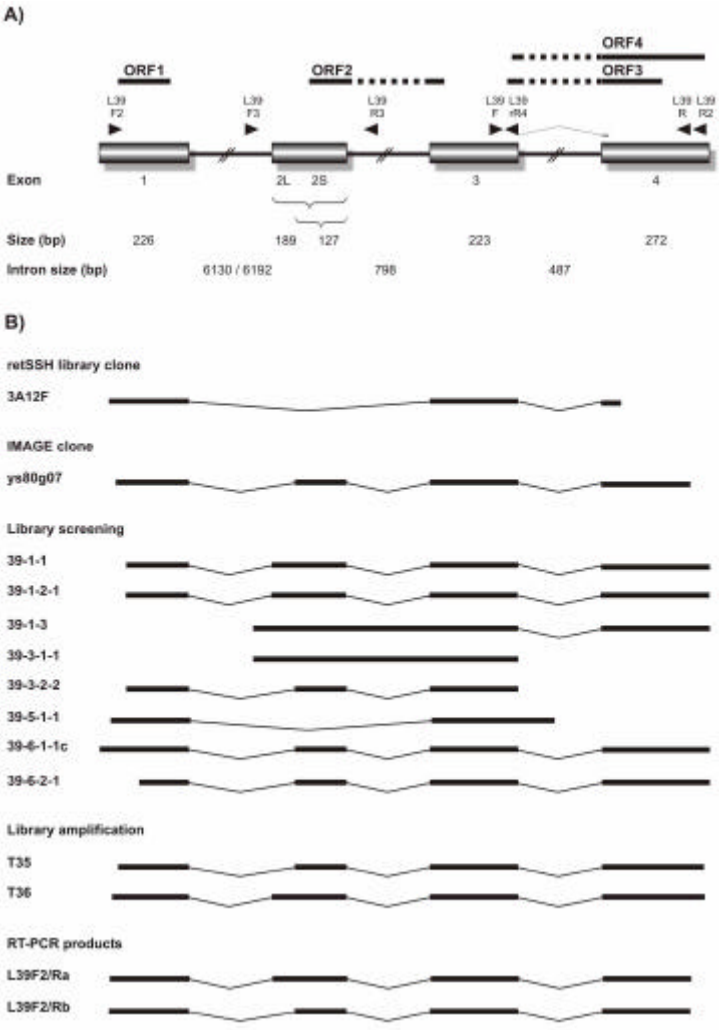


Fig. 49 Schematic representation of C4orf11

A) The gene structure is depicted by boxes representing the exons (proportional to original size) and lines representing the introns (not proportional). The primers used in the cloning process are indicated by arrows above the diagram. The cDNA contains four different ORFs which are indicated by the lines above the schema.

B) Sequences of various sources were used to assemble the cDNA sequence of the gene. The source, lab intern ID, and genomic organization of each sequence is indicated.

In an effort to prolong the 5'-end of the gene, an aliquot of the retina library constructed by Jelena Stojic was amplified using primers L39R2 and ?TriplEx5' (Table 35). Two strong products (of approximately 840 and 900 bp) could be amplified and were cloned (T35 and T36). Sequencing of the products verified the already known sequences but did not extend the 5'-end. The products, which differed in size due to the use of the alternative splice site in exon 2, ended 29 bp downstream of the 5'-end of clone 39-6-1-1c.

Two cDNA sequences which include only the sequences which were found in both the PCRs and the clones were assembled. Isoform C4orf11_v1 (GenBank acc. no. AY316301) is 910 bp and includes the longer exon 2. C4orf11_v2 (GenBank acc. no. AY316302) is 848 bp long. A polyadenylation signal is found just 20 bp upstream of the 3'-end suggesting that the complete 3'-end of the gene has been identified. There is a great content of repeat sequences in this gene since more than 50% of the sequence derives from an inserted LTR sequence. Both ends of the gene are composed exclusively of LTR-repeat sequences.

6.6.2. Genomic structure of C4orf11

The 0.9 kb sequence of C4orf11_v1 is encoded by four exons (Table 31) distributed over 8.3 kb on chromosome 4 within band q21.22. Since two different splice acceptor sites are found in exon 2, there are at least two different variants of the gene, which differ by a 62 bp insertion. Intronic sequence may be retained in some transcripts, as supported by cDNA library clones 39-1-3, 39-3-1-1, and 39-5-1-1. Retention of the IVS1 and 2 would not alter the open reading frame, but inclusion of IVS3 would result in a premature stop.

A comparison of the transcripts and the genomic sequence revealed six base changes in the C4orf11_v1 sequence (c.-451A>G, c.-287G>G, c.-148T>C, c.82A>G, c.162A>C, and c.266T>C). Since the frequency has not been investigated in the general population, it is not possible to assert which ones might constitute true polymorphic SNPs.

Table 31 Exon / intron structure of C4orf11

Exon		3'-Acceptor Splice Site[a]		5'-Donor Splice Site[a]		Intron	
No.	**Size (bp)**	**Sequence**	**Score[b]**	**Sequence**	**Score[b]**	**No.**	**Size (bp)**
1	226			CAGgtaagt	0.00	1>2L	6130
						1>2S	6192
2L	189	tgattgtaaccaacagG	9.44	GGGgtgagc	3.20	2	798
2S	127	ggatgatgatattcagG	11.88	GGGgtgagc	3.20		
3	223	ttgatttctcctatagG	4.48	CAGgtattc	6.14	3	487
4	272	cttcttcccctgccagG	2.85				

[a] Exonic and intronic sequences in upper and lower case letters, respectively.
[b] Score of donor/acceptor splice site. According to published data (Berg and von Hippel 1998 and Penotti 1991) 99% of sites have a score of 0-11 (donor) or 0-20 (acceptor). Scores were calculated using the spreadsheet created by Christian Sauer, 2001.

6.6.3. Expression analysis of C4orf11

The expression of C4orf11 was determined by RT-PCR (Fig. 50A), by virtual Northern blot (Fig. 50B), and by qRT-PCR (Fig. 50C). The RT-PCR using primers L39F/L39R (Fig. 49 and Table 36) amplified exclusively a product from the retina cDNA (Fig. 50A). For virtual Northern blot analysis, filter DS14 was incubated with a radio-labeled probe amplified with primers L39F and L39R (Table 36). Two very strong hybridization signals of approximately 0.85 and 0.80 kb were seen in retina, a weak hybridization could be verified for RPE (Fig. 50B). The qRT-PCR with primers L39rR4/L39R (Table 37) confirmed the retinal/RPE expression, but based on the relative quantities obtained by this method, the gene may be expressed at higher levels in the RPE (Fig. 50C).

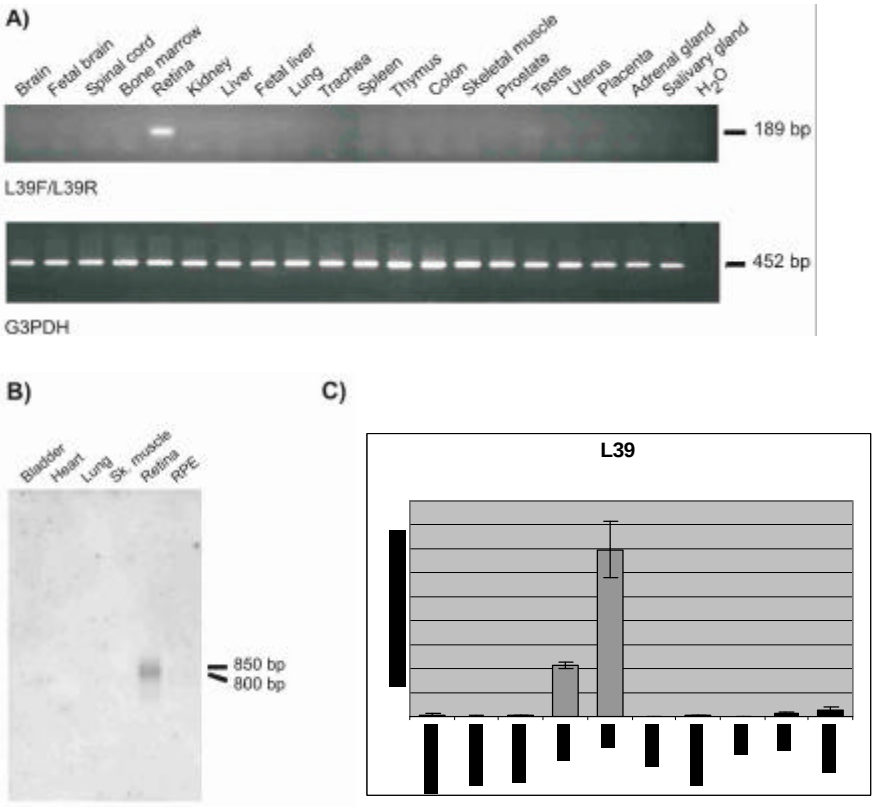


Fig. 50 Expression analysis of C4orf11
A) Expression analysis by RT-PCR using primers L39F/L39R determined that the gene is expressed at high levels only in retina; minor expression could be seen in testis.
B) The transcript size of C4orf11 could be determined by virtual Northern blot using the L39F/L39R amplicon as a probe. The retina expression was confirmed but faint bands were also seen in RPE. Apparently there are two or more transcripts; this correlates well with the various sequences found in the retina library.
C) Quantification of gene expression by qRT-PCR using primers L39rR4/L39R it suggests that the gene is expressed at higher levels in RPE than retina. The expression measured in basal ganglia, distal colon, lung, and stomach is fictitious and due to unspecific products.

6.6.4. *In-silico* analysis of the putative C4orf11 proteins

Translation of both isoforms of C4orf11 results in the same ORFs. There are four different hypothetical ORFs encoding 44, 47, 70, or 94 aa, respectively (Fig. 49). The first ORF (47 aa, ORF1) is encoded entirely in exon 1. The 44 aa ORF (ORF2) is encoded partially by exons 2 and 3. Finally, the ORFs with 70 (ORF3) and 94 aa (ORF4) both start in exon 3 and extend to exon 4, but represent two reading frames.

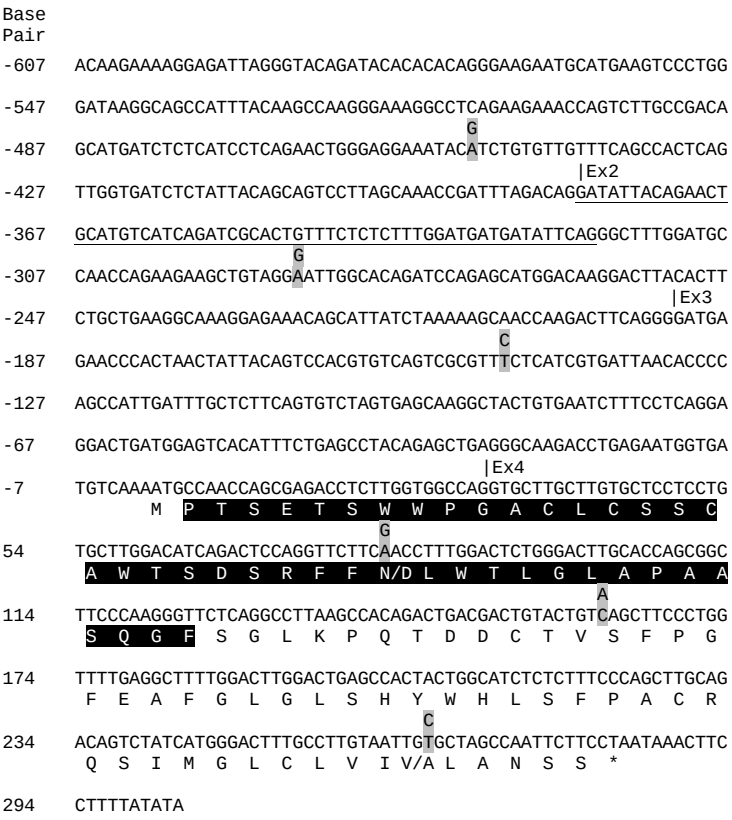


Fig. 51 cDNA sequence and putative C4orf11_ORF4 protein
The underlined sequence is included in C4orf11_v1, but is spliced-out in variant 2. Exon boundaries are indicated by vertical lines. The gray-shaded bases show the location of the base changes found. The putative polyadenylation signal is underlined and grey shaded. The amino acid sequence shown in this figure corresponds to ORF4, one of the four hypothetical proteins that are encoded by the gene. Its putative signal peptide is indicated by a black background. In the cases where the base change leads to an amino acid exchange, both amino acids are shown separated by a slash. The numbering on the left corresponds to the nucleotides, the one on the right to the amino acids. The exact exon boundaries are depicted by a vertical line followed by the abbreviation 'Ex' and the corresponding exon number.

To determine the true ORF of the gene, the properties and homologies of all ORFs were investigated. Searches in the NCBI databases with BLASTP[73] only identified protein similarities to C4orf11_ORF4. The highest homology was with XP_209376, a hypothetical protein predicted from a testis full-length clone (AK093119) mapping to chromosome 21q21.3. Over 76 aa, sequence identity is 43% and similarity is 54%. Interestingly, this 101 aa protein is encoded by a 5-exon gene with the first three exons being non-coding. Although the ESTs containing sequence from XP_209371 derive from testis and colon, there are several retina ESTs mapping to the intron sequence of the gene. There is also a hypothalamus clone which contains the first two exons, but then continues with sequence from IVS2. Therefore, the general structure of the XP_209371 locus greatly resembles the C4orf11 organization.

The C4orf11_ORF4 is also similar to the N-terminal region of a hypothetical protein translated from the stomach full-length clone AK097340 located on chromosome 8q21.11. This gene consists of five exons, the first of which seems to be alternatively spliced. Due to the alternative exon usage, the hypothetical proteins are different. The homology of C4orf11_ORF4 is only to a 55 aa fragment of the hypothetical protein translated from the stomach clone. More than 60% of the amino acids are chemically similar.

Since the other ORFs are shorter and are not similar to other proteins, only the predictions for C4orf11_ORF4 will be considered. *Ab initio* predictions of protein function by ProtFun[74] assign the putative C4orf11_ORF4 protein to the growth factor gene ontology group and suggest that it may be

[73] http://www.ncbi.nlm.nih.gov/BLAST/
[74] http://www.cbs.dtu.dk/services/ProtFun/

an enzyme involved in energy metabolism. It is predicted to be a glycosylphosphatidylinositol-anchored protein and to contain a signal peptide. These predictions are in contradiction with PSPORT[75], which locates the C4orf11_ORF4 in the cytoplasm. The supposition that the protein may be extracellular is supported by the glycosylation prediction of threonine 11.

6.7. Cloning and characterization of the death associated protein-like 1 gene (DAPL1)

The characterization of DAPL1 began after sequencing of clone 5G2F from the retSSH library. Alignment of the 0.55 kb 5G2F sequence to the human genome sequence using the USCS Genome Browser[76] revealed a number of other ESTs which partially or totally overlapped with 5G2F. They were grouped in the Hs.59761, which contains 27 sequences including a clone derived from an eye library (UI-E-EJ1) and three from the pineal gland II library. The 5G2F clone contained all but the last eight bases of the consensus 552 bp cDNA.

The DAPL1 gene is organized in four exons which cover 20.7 kb of genomic sequence on chromosome 2q24.1 (Table 32). The sequence of DAPL1 has been submitted and has been assigned GenBank accession number AY324399.

Table 32 Exon / intron structure of DAPL1

Exon		3'-Acceptor Splice Site[a]		5'-Donor Splice Site[a]		Intron	
No.	Size (bp)	Sequence	Score[b]	Sequence	Score[b]	No.	Size (bp)
1	114			CAGgtaggc	2.82	1	8851
2	88	ttttatctgtttgtagT	6.61	AAGgtaggg	2.92	2	2685
3	61	cttttcatcttcacagT	4.42	AAGgtgagc	1.23	3	8589
4	289	cccttctcacctttagC	5.57				

[a] Exonic and intronic sequences in upper and lower case letters, respectively.
[b] Score of donor/acceptor splice site. According to published data (Berg and von Hippel 1998 and Penotti 1991) 99% of sites have a score of 0-11 (donor) or 0-20 (acceptor). Scores were calculated using the spreadsheet created by Christian Sauer, 2001.

6.7.1. Expression analysis of DAPL1

Expression analysis with primers L93F/L93R (Table 36) in a panel of nine different cDNAs established that DAPL1 is expressed at high levels in retina, RPE, thymus, and brain (Fig. 52A). To determine the size of the transcript, a virtual Northern filter (DS16) was hybridized with a probe amplified with primers L93F/L93R. A 0.6 kb transcript was present only in retina and RPE, with strongest expression in the latter (Fig. 52B). No evidence for additional transcripts could be detected. Therefore, this result provided evidence that the full-length sequence of the gene had been cloned. The expression was confirmed by qRT-PCR with primers L93F2 and L93R2. The expression in retina and RPE is 332 and 527 fold higher, respectively, in relation to the expression in heart (Fig. 52C).

[75] http://psort.nibb.ac.jp/form2.html
[76] http://genome.ucsc.edu/cgi-bin/hgBlat

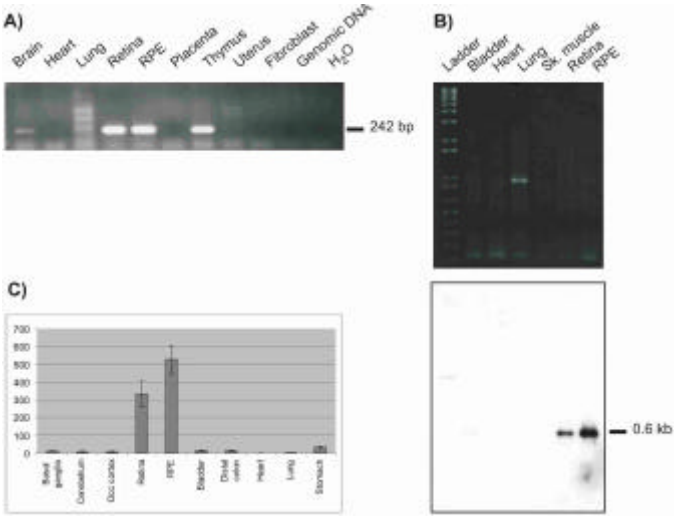


Fig. 52 Expression analyses of DAPL1
A) The expression of DAPL1 was originally tested with primers L93F/L93R in a panel of nine tissues plus a genomic DNA and H_2O controls. The 242 bp product was amplified in significant quantities in retina, RPE, and thymus. Less expression was observed in brain, heart, lung, and uterus.
B) The expression obtained by RT-PCR was corroborated by virtual Northern blot. The upper picture was taken after electrophoretical separation of the full-length double-stranded cDNAs. After hybridization with the radio-labeled probe, amplified with primers L93F/L93R from retina, a very strong signal of approximately 0.6 kb was seen in RPE and a somewhat weaker signal in retina. No expression was seen in any other tissue.
C) A qRT-PCR amplification with primers L93F2 and L93R2 confirmed that the gene is expressed at significant levels only in retina and RPE.

6.7.2. *In-silico* analysis of the putative DAPL1 protein

The 552 bp mRNA transcript contains an ORF encoding a 107 aa putative protein (Fig. 53). None of the representative sequences from the gene, including the one cloned in the present study, include the in-frame stop codon which is present 52 bp upstream in the genomic sequence. There are six possible donor acceptor sites (AG) in this fragment, but it appears unlikely that these correspond to true splice consensus sequences as the calculated scores range from 15.4 to 18.8 whereas scores regularly lie between 0 and 14. The protein, which has a calculated molecular weight of 11.9 kDa and an isoelectric point of 10.0, contains no Pfam motif, domains, or signal peptide. It does contain an ER membrane domain (QPRK) and is predicted to localise in the nucleus (82% probability).

```
      Exon1
-56   ACAGCTGGCATTCAGCCTCCAGAGCACCAGCACTGGCACTGGCACTGGCACACGCTATGG
                                                              M  A   2
                                                                |Exon2
5     CAAATGAAGTGCAAGACCTGCTCTCCCCTCGGAAAGGGGGACATCCTCCTGCAGTAAAAG
        N  E  V  Q  D  L  L  S  P  R  K  G  G  H  P  P  A  V  K  A  22

65    CTGGAGGAATGAGAATTTCCAAAAAACAAGAAATTGGCACCTTGGAAAGACATACCAAAA
        G  G  M  R  I  S  K  K  Q  E  I  G  T  L  E  R  H  T  K  K  42
                           |Exon3
125   AAACAGGATTCGAGAAAACAAGTGCCATTGCAAATGTTGCCAAAATACAGACACTGGATG
        T  G  F  E  K  T  S  A  I  A  N  V  A  K  I  Q  T  L  D  A  62
                              |Exon4
185   CCCTGAATGACGCACTGGAGAAGCTCAACTATAAATTTCCAGCAACAGTGCACATGGCGC
        L  N  D  A  L  E  K  L  N  Y  K  F  P  A  T  V  H  M  A  H  82

245   ATCAAAAACCCACACCTGCTCTGGAAAAGGTTGTTCCACTGAAAAGGATCTACATTATTC
        Q  K  P  T  P  A  L  E  K  V  V  P  L  K  R  I  Y  I  I  Q  102

285   AGCAGCCTCGAAAATGTTAAGCCTGGATTTAAAACACAGCCGTCTGGCCAGCTGCCTCGA
        Q  P  R  K  C  *                                          107

345   ATATCTGACAGCTTAGCAAAAAGGGCCAAAGCTTTCCATAGGCGTGCTGCACTTGCTTGG
385   TAAATTAAGCAGCTTTTGTATCTTCCCCTTTGACTTTAGGTAATAAAGCATCCAAACTTG
445   TAAATCTGACAC
```

Fig. 53 cDNA and protein sequence of DAPL1
The 512 bp of L39 are organized in four exons with the start codon (bold) located in the first and the stop codon (bold) in the last exon. Vertical lines above the sequence indicate boundaries. The polyadenylation signal (underlined) is found just 35 bp upstream of the end of the sequence. The 107 aa putative protein (numbering on the right) contains an ER membrane domain (grey background).

Homology searches identified a number of homologues (Fig. 54). The highest homology, 85%, was with a murine protein from the RIKEN collection (XP_130301). Similarity with a rat prediction (XP_229942) was 82%. Other sequences with high similarity include: zebrafish DAP1b (NP_571648) and DAP1a (NP_571647), human DAP (NP_004385), and the Longjawed mudsucker DAP

(AAL16035). Proteins which have not been annotated but also probably belong to the same family include NP_666169 (*Mus musculus*), NP_610676 (*Drosophila melanogaster*), and AAH45071 (*Xenopus laevis*). A BLASTN search[77] in the EST and nr database identified three additional sequences with high homology, namely BM433753 (*Bos Taurus*), BI181845 (*Sus scrofa*), CB559023 (*Xenopus laevis*), and AI111671 (*Rattus norvegicus*). These sequences were translated and also included in a sequence alignment (Fig. 54).

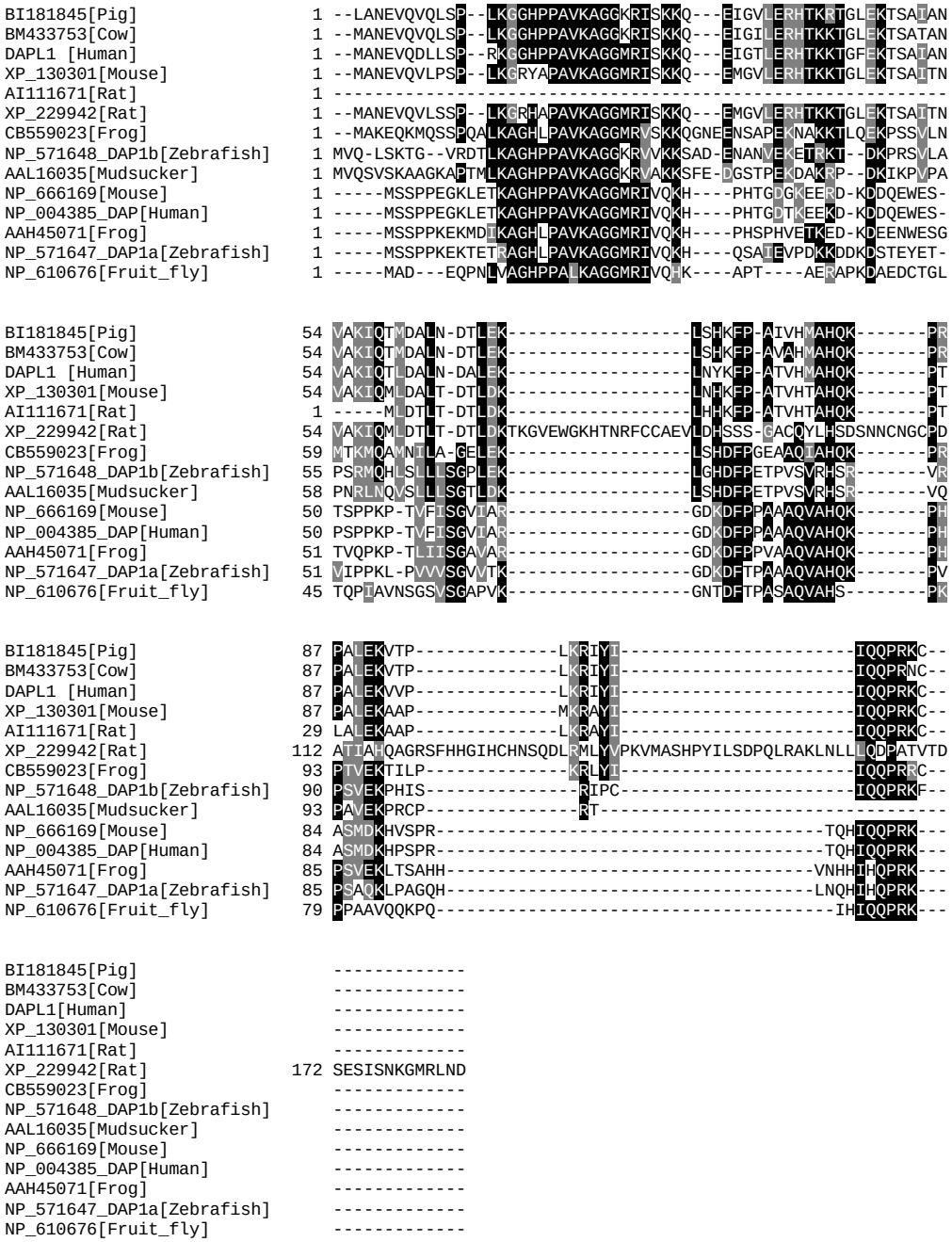


Fig. 54 Multiple sequence alignment of the human DAPL1 and related proteins
Proteins and hypothetical proteins from nine different species were aligned. In some cases, the protein has been obtained by translation of clones which are not full-length (e.g. AI111671) or from predictions based on genomic sequence (e.g. XP_229942). Therefore, the sequences should not be viewed as definite, but are shown just to illustrate the high homology and conservation of certain motifs between the family members. In the C-terminal an ER membrane retention signal (QPRK) is observed.

[77] http://www.ncbi.nlm.nih.gov/BLAST/

As mentioned above, some of the related proteins are already identified as the death-associated proteins (DAP). The human DAP gene maps to chromosome 5p15.2 and encodes a basic, proline-rich, 15 kDa protein whose name derives from the fact that this protein and the death-associated protein kinase-1 are indispensable for the execution of programmed cell death (Feinstein et al. 1995). The death-associated protein acts as a positive mediator of programmed cell death that is induced by interferon-gamma. From the chromosomal localization it is obvious that DAPL1 is not identical to DAP.

A phenogram was constructed from all the aligned sequences (Fig. 55). The alignments show that DAPL1 belongs to a family with at least three additional members: DAP, DAP1a, and DAP1b. The phylogenetic classification suggests that the pig, cow, and rat sequences plus the XP_130301 mouse protein may be the orthologous genes of DAPL1. The human DAP protein seems more related to the DAP1a zebra fish than to DAP1b.

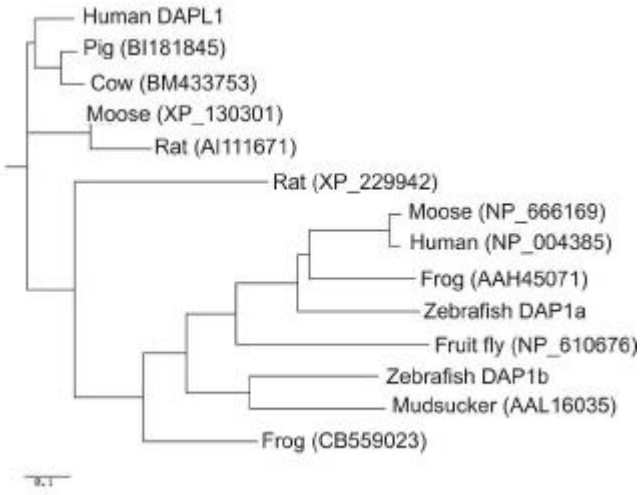


Fig. 55 Phylogenetic analysis of DAPL1 and related proteins
The proteins of various species which share sequence identity with DAPL1 were aligned and compared. The phylogenetic tree of the results is shown on the left. As can be seen DAPL1 is more closely related to the pig and cow proteins BI181845 and BM433753 than to DAP1a, DAP1b (zebrafish), or DAP (human). The ruler in the bottom left corner gives an idea about the distance between the proteins.

The high homology between the various species and different DAP proteins and DAPL1 is astonishing considering the fact that a Pfam analysis reported that the DAPL1 protein does not belong to any recognized protein family. There might be functional differences between the proteins as reflected by the fact that the ProtFun[78] prediction vaticinates that DAP has a regulatory function involved in transcription whereas a role in translation and as a growth factor is anticipated for DAPL1.

[78] http://www.cbs.dtu.dk/services/ProtFun/

VI Discussion

The unique histological and metabolic characteristics of the retina are due to its distinctive gene expression profile. A survey of the studies published until the end of the 1990s revealed that a number of retina genes had been identified, but principally as a result of phenotype-based approaches. To complement these efforts and aid the identification of novel retinal genes, the present project was initiated to systematically characterize the retinal transcriptome applying a gene-based approach. Since the analysis of the expression profile of the known genes associated with retinal degenerations disclosed that a majority are preferentially expressed in the retina, our efforts concentrated on the generation of a catalogue of genes expressed exclusively or preferentially in the retina or in both the retina and the central nervous system. The challenges encountered in the data-mining process and the strategies applied to establish the expression profile of these genes will constitute the first part of the discussion. In the second part, the characteristics of selected novel genes will be discussed.

1. Deciphering the retinal transcriptome

The ultimate purpose of the human genome project was not to simply provide the scientific community with three billion sorted bases, but to enable scientists to make sense of this sequence by providing the basis so that all human genes and their encoded proteins could be compiled. Thus, the long-term aim of the human genome effort is the compilation of all genes in a 'Periodic Table of Life' which would enable the understanding not only of particular components but of whole biological systems (Lander 1996, Peltonen and McKusick 2001).

Gene identification in lower organisms is almost trivial because the absence of introns in bacteria and their paucity in yeast means that most genes can be readily recognized by *ab initio* analysis (Lander et al. 2001). In higher organisms such as fly and worm, coding sequences comprise a large proportion of the genome and thus, even though they contain introns, gene identification is not particularly difficult. Since only 1% of the human genome is spanned by exons, whereas 24% is intronic, and 75% of the genome is intergenic DNA (Lander et al. 2001), the large-scale identification and characterization of novel genes only became possible with the availability of the genome sequence and novel bioinformatic technologies. Currently, programs can correctly identify up to 75% of exons but the rate of false-positives is still very high and less than 50% of predicted gene structures correspond to actual genes (Rogic et al. 2002). Particularly initial and terminal exons tend to be the most difficult to identify (Stormo 2000). Therefore, the current limitations of *a priori* genome annotation dictate that human transcripts still need to be characterized experimentally.

The distinct functions fulfilled by the retina require a great degree of differentiation which is achieved by the presence of more than 50 different cell types (Masland 2001). This diversity of cell types results in the existence of a large complement of transcripts and it has been suggested that approximately 25,000 to 27,000 genes could be expressed in the mammalian retina (Swaroop and Zack 2002, Mu et al. 2001, respectively). In the last couple of years, this realization and the fact that the retina has been a forerunner in neuronal research (Dowling 1987), has led to multiple efforts centered on the compilation of a list of genes expressed in the human, murine, bovine, and canine retina (Table 33). The techniques used to characterize the retinal transcriptome have included *in-silico* projects (Sohocki et al. 1999, Malone et al. 1999, Bortoluzzi et al. 2000, and Katsanis et al. 2002), sequencing of cDNA libraries (Shimizu-Matsumot et al. 1997, den Hollander et al. 1999, Sinha et al. 2000, Mu et al. 2001, Lin and Sargan 2001, Sharma et al. 2002, and Wistow et al. 2002), serial analysis of gene expression (SAGE) (Blackshaw et al. 2001, Blackshaw et al. 2003, Sharon et al. 2002), and cDNA microarray analysis (Hackman et al. 2003 and Chowers et al. 2003).

Table 33 Survey of transcriptome studies of the mammalian retinal transcriptome by other groups

	Methodological approach	No. of genes reported[a]			No. of predicted retina-specific genes (of total)	Reference
		known	novel	total		
***In-silico* projects**	TIGR clusters (human retina and pineal gland)	12	33	45	25	Sohocki et al. 1999
	UniGene clusters (human)	2284	2690	4974	536	Bortoluzzi et al. 2000
	dbEST (human)	n.a.	n.a.	925	30[c]	Katsanis et al. 2002
cDNA library sequencing projects	Substracted cDNA library (human retina)	n.a.	n.a.	607	43	Shimizu-Matsumoto et al. 1997
	Subtracted cDNA library (human retina and RPE)	341	99	440	33[c]	den Hollander et al. 1999
	Substracted cDNA library (human retina)	35	23	58	n.a.	Sinha et al. 2000
	cDNA library (mouse retina)	1745	3323	5068	n.a.	Mu et al. 2001
	cDNA library (canine retina)	29	57	100	n.a.	Lin and Sargan 2001
	cDNA library, unamplified and un-normalized (human retina)	n.a.	n.a.	1254	n.a.	Wistow et al. 2002
SAGE projects	SAGE (mouse)	187	75	262	n.a.	Blackshaw et al. 2001
	SAGE (human)	690	47	737	89	Sharon et al. 2002
cDNA library microarray	cDNA library microarray and *in-situ* hybridization of candidates (embryonic chick retina)	n.a.	n.a.	5000	272	Hackam et al. 2003
	I*n-silico* analysis and cDNA microarray (human)	113	98	211	84	Chowers et al. 2003

n.a.: not available
[a] The numbers given are corrected for redundancy
[b] Total number includes all non-redundant TIGR clusters containing more than three sequences
[c] Expression was experimentally verified

Based on the available data (Table 33), it should be feasible to assemble a first-draft transcriptome. In practice this is not possible yet since the majority of the publications include only a selection of the identified genes which makes the compilation of comprehensive list impractible. Nevertheless, from the available information it is apparent that there is no significant overlap between the results obtained in each study. For example, from the 180 Hs. clusters investigated in Phase I of our UniGene project, none were included in the set of 203 genes identified in the Blackshaw study (2001), only two were found by Sharon et al. (2002), and 28 were found in the set of 284 genes reported by Chowers et al. (2003). A similar trend is observed when the 321 known genes sequenced from our retSSH library are compared with the 2114 known genes sequenced from the retNEIBank collection (Wistow et al. 2002). Only 122 genes were sequenced from both libraries, with 199 of the genes from the retSSH library being unique.

Although some of the aforementioned studies have investigated a large number of genes and reported many candidate genes, except for the den Hollander et al. (1999) and the Katsanis et al. (2002) studies, the expression of the identified candidates has not been verified. Therefore, the approach used in our project is unique because not only thousands of sequences were analyzed, but the expression of hundreds of them was systematically investigated in many tissues by a number of different methods.

Among the methods applied to characterize transcriptomes, sequencing of ESTs has been particularly informative (Burge 2001). The idea of EST sequencing, developed and published by Adams et al. in 1991, quickly became a widely used method of gene discovery. In 1992, Okubo et al. proposed the use of ESTs as a way of estimating the level of gene expression based on the premise that the number of ESTs that represent a gene gives a rough indication of the expression level of the gene in the tissue. The advantage of this method is that it is not necessary to know *a priori* the sequence of the transcripts. On the other hand the sequences are long enough to be able to identify, if desired, the full-length sequence.

Although the rate of novel gene discovery by the EST method has slowly declined (Martin and Pardee 2000), it is still very useful to characterize gene expression. Part of the success is due to the suppression subtracted hybridization (SSH) technique which was introduced to generate libraries normalized and enriched for particular transcripts (Duguid and Dinauer 1990, Diatchenko et al. 1996). Generating cDNA libraries with this method has the advantage that high subtraction efficiency with an equalized representation of the transcript population can be achieved. For this reason, it is a very valuable tool for the discovery of the estimated 15,000 genes expressed at levels of less than 5 copies per cell (Bonaldo et al. 1996).

The contribution of this project to compile the retinal transcriptome was based on two complementary EST-based approaches. First, the ESTs deposited in the public gene indexing database UniGene were assessed and based on the retinal-EST content clusters were chosen for *in-silico* and expression analyses. Although this pursuit is adequate for the identification of genes expressed at high levels it is

not very useful to identify rare genes. Therefore, it was complemented with a second approach geared towards the identification of rare retina-specific genes which was accomplished by the generation and random sequencing of clones from a retina SSH (retSSH) cDNA library.

2. Evaluation of the UniGene approach

The UniGene approach was conceived to mine the publicly available EST data in order to identify genes expressed preferentially in the retina in the most efficient way. At the time of analysis (June 2000), 17,736 retina ESTs were grouped in 6190 clusters. From these, 2201 contained the sequences of already known genes and were therefore removed from further analyses.

The first stage of the UniGene approach was designed to investigate the power of cluster composition to predict the expression of a gene. For this purpose, the 1241 Hs. clusters representing unknown genes which had a retina-EST fraction of at least 30% were selected. After *in-silico* analysis the clusters were grouped in two categories. More than half of these clusters (673) were composed exclusively of retina ESTs and were included in category A. Category B was reserved for 568 clusters which also included ESTs from other tissues. To increase the power of analysis, each category was subdivided in four groups containing increasing number of retina ESTs per cluster as this may be an indication to predict retinal expression. The research hypothesis were that either category A would contain a bigger fraction of genes preferentially expressed in retina or those clusters which contained greater number of retina ESTs would account for a preferential expression. To test this, 180 clusters representing unknown genes were selected from all subcategories and their expression was determined by RT-PCR. A total of 39 clusters were found to be expressed specifically in retina and 32 presented neuronal expression. These results reveal that the percentage of tested genes with retina expression is of 21% for category A and 23% for category B (Fig. 56). The same trend (17 and 18%) applies to the neuronal expression. Based on this, it can be concluded that clusters containing only retina ESTs are not more likely to be preferentially expressed in retina than those containing ESTs from other tissues as long as the proportion is smaller than 70%.

There were, however, some differences between the two groups. Whereas a significantly higher number of clusters (25%) from category A were not expressed in any of the tested tissues (Fig. 56), the number of ubiquitously expressed genes from category B is higher (49 versus 37%). The second hypothesis does not seem to be applicable either since within the subcategories A and B, the proportion of EST clusters exhibiting different types of in vitro expression patterns does not vary substantially as the number of retina ESTs increases (Table 11).

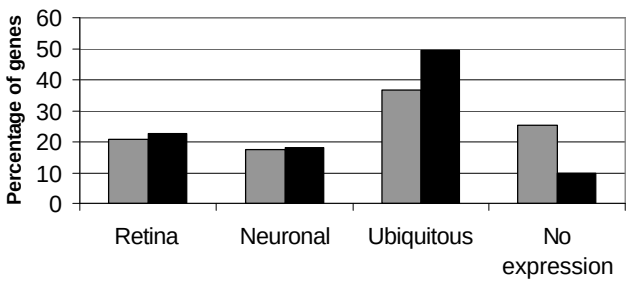


Fig. 56 Expression of clusters from categories A and B of the Phase I UniGene analysis
There is no significant difference in percentage of genes with retina and neuronal expression between clusters which contained exclusively retina ESTs (category A - grey bars) and those that also contained ESTs from other tissue (category B - black bars). A difference between the number of ubiquitously expressed genes and products that are not amplified in any of the tissues could be observed.

The results from the pilot study (Phase I) proved that there is no direct correlation between a high level of retina EST in a cluster and a retina-specific expression. Therefore, for the second phase of the UniGene data mining, the selection of genes for expression analysis was not based on cluster composition but on bioinformatical analyses.

To achieve this, the remaining 293 Hs. clusters with at least two ESTs and a retina-EST proportion of at least 30% at the time of the initial selection were analyzed in detail using bioinformatic tools. The criteria applied to filter and catalogue the clusters included the survey of homology of the sequence to other species, evidence of splicing, absence of repeats, and overlap to gene predictions. From the 73 clusters selected for expression analysis by RT-PCR, five (7%) were expressed exclusively in retina and seven (10%) in neuronal tissues. Fourteen percent of the primer pairs failed to amplify any of the tested cDNAs. The 'no expression' fraction is lower than the proportion observed for category A, but higher than the one of category B of the Phase I analysis. Therefore, in spite of the *in-silico* efforts to predict good candidates, the percentage of retina and neuronal genes identified in Phase II was less successful than in the Phase I approach.

The UniGene data-mining strategy was useful to identify 44 retina-specific genes and 39 genes expressed in the neuronal system. From the 44 retina-specific genes, nine have been cloned. Seven genes have been cloned in our group (C12orf3, C12orf7, C1orf32, F379, MPP4, NETO1, and RFRP), SIX3 has been cloned by Granadino et al. (1999), and one gene has been sequenced in the NEDO human cDNA sequencing project but has not been published. Of the 39 neuronal genes, 16 have been cloned, the sequence of one is part of a prediction, and three are probably splice variants of already-known genes. Three of the neuronal genes (C14orf29, GRM7, and WDR17) were cloned in our laboratory.

Close evaluation of our results demonstrates that the UniGene approach suffers from several limitations possibly shared by all EST-based analyses. One obstacle lies in the fact that the representation of a transcript in the EST collection depends on the number of ESTs that have been sequenced and the quality of the libraries. An undoubtedly important factor for the study of the retina transcriptome is the fact that at the time the UniGene collection was mined more than half of the retina ESTs had been sequenced from a single library, the Soares retina library N2b5HR. This certainly led to the existence of many artefacts in the data set, possibly as a result of contamination with unspliced

mRNA (Croft et al. 2000) or genomic DNA. Approximately 44% of the sequences from the Phase I analysis map within introns of known genes (Table 10). In a study by Sorek and Safer (2003) the proportion of intron-derived sequence present in 1906 libraries was investigated and they found that the mean percentage of intron sequences was 3.7% with a standard deviation of 2.0%. The authors conclude that if more than 9.8% of the sequences from a library derive from intronic sequence, the library should be considered to be highly contaminated with pre-mRNA sequences. In our case, we are not considering a single library but assemblies of various libraries and we have omitted the known genes therefore it is not possible to transfer this cut-off as such. Nevertheless, the percentage of sequences mapping within the intron of an already known gene is seemingly high. As a reference, in the retSSH library only 18 from 87 clusters (20%) mapped to the intron of a gene.

3. Evaluation of the retSSH approach

The second method used during this project was the analysis of sequences isolated from the retSSH library. This library was generated by subtracting from a retina sample those transcripts also present in liver and kidney. From 1093 randomly picked clones, 13 derived either from mitochondria or contained only vector sequence, therefore a pool of 1080 clones was available for analysis. After adjustment for redundancy 413 clusters could be assembled. From the 413 transcripts, 248 were known genes in May 2001 and this number increased to 321 two year later (April 2003 assembly).

The frequency of anonymous clones, i.e. those that are not similar to either known genes or annotated sequences, of this collection falls within the range of 40% anonymous ESTs found by other groups (Lee et al. 2002). Since the draft human genome sequence was already available at the time when the clustering was done, the assembly of the anonymous clones was achieved by comparison to the human genome sequence. The percentage of singletons from the known gene set (68%) is not significantly lower than for the unknown gene set (76%), therefore it is likely that most clones that derive from the same gene have been properly assembled.

The expression profiling of the retSSH collection included all clusters representing unknown genes. It must be noted that several of these genes were already reported full-length sequences, but they were included in the expression analysis since their expression had not been investigated so far and they had not received official gene names. Expression profiling of 89 clusters, revealed that 25 (28%) are expressed specifically in retina and 12 (13%) in the nervous system. The remaining 52 (59%) genes are ubiquitously expressed.

To date, 17 of the 25 retinal genes have been cloned and the full-length sequence of nine of the 12 neuronal genes is known. Our group has worked on the characterization of a number of these genes leading to some interesting findings which include for example the identification of three RNA genes. Two of the genes (C4orf11 and DAPL1) cloned as part of this thesis were originally identified in the retSSH library.

The quality of the retSSH library was assessed by in-depth analyses which included comparison to the publicly available retina cDNA library (retNEIBank) and to different collections of genes. The comparison to retNEIBank was not only useful to evaluate the degree of subtraction and suppression attained in retSSH but also to estimate the overlap that may exist between various gene-identification approaches for human retina. From the 2615 unique transcripts of the retNEIBank library, 2114 were known genes. Only 122 of them were also found in the retSSH collection of known genes (122 out of 321). The comparison highlights the value of a normalized subtracted library as two-thirds of the retSSH genes had not been yet found in the six times larger retNEIBank collection. Therefore these genes are probably expressed at low levels in the cell.

As a result of the different construction methods, the genes sequenced more frequently in each library varied greatly. Only three genes (RHO, SAG, and GLUL) were found frequently in both libraries. A careful analysis highlights another advantage of subtracting common transcripts. Whereas only 16% of the most abundant genes in the retSSH library do not have a retina or neuronal-restricted expression, 50% of the common genes of retNEIBank are ubiquitously expressed. The subtraction efficiency was also analyzed by comparing known genes of each collection with housekeeping genes extracted from HuGEIndex Gene Specific Expression database[44]. In this case 17.8% of the known genes of the retNEIBank collection are ubiquitously expressed in comparison with only 9.9% of the ones from the retSSH library.

Another parameter that can be used to assess subtraction efficiency is the fraction of ribosomal protein genes (Bortoluzzi et al. 2001) found in the libraries. Whereas 59 (70%) of the ribosomal protein genes have already been sequenced from the retNEIBank collection only one has been found in the retSSH library. An interesting observation is that based on the *in-silico* data used by Bortoluzzi et al. (2001), 19 of the 84 ribosomal genes were not expected to be expressed in retina. Contrary to this assumption, the retNEIBank library contains transcripts corresponding to nine of them (RPL6, RPL18, RPL22, RPL26, RPS2, RPS3, RPS8, RPS16, RPS17).

A final indicator of the value of normalization and subtraction was obtained by comparing genes known to be vital for proper retinal function. In the retSSH collection, 19 (36%) of the 53 genes that are involved in phototransduction and the vitamin A cycle were represented by at least one clone (Table 16). From the 2114 known genes sequenced from the retNEIBank library, 26 (26/53, 49%) are involved in phototransduction or the vitamin A cycle (Table 16). Therefore, even though the retNEIBank collection is six times larger as the retSSH, the number of genes from these pathways identified in the retNEIBank collection was only 13% greater. From 194 genes possibly involved in synaptic transmission, only 8/194 (4%) have been sequenced from the retSSH library (Table 17). This information is a rough estimation because some of the genes included in the list may be exclusively used in brain synapses and therefore will never be found in retina. Nevertheless, the fact that the retNEIBank library has been more extensively sequenced as the retSSH is reflected by the fact that 33 of these 194 genes (17%) have already been sequenced from this collection (Table 17).

[44] http://www.hugeindex.org/

As can be seen, the estimation of the number of genes expressed in a cell is a challenging biological problem (Bishop et al. 1974). Theoretically, by exhaustive sequencing of a SSH library, it would be possible to determine the number of tissue-specific genes. To see if it would be possible to extrapolate results and estimate the number of tissue-specific genes at the present stage of sequencing of the retSSH library, varied statistical analyses of the 1080 clones of the retSSH library were attempted. Even though different models, described in Kuznetsov et al. (2002), Bunge and Fitzpatrick (1993), and Stern et al. (2003), were adopted for the estimation, no reliable and conclusive results could be obtained. A linear regression plotting of the number of sequenced clones versus the number of novel genes determined that the rate of discovery of novel genes is still in the linear range, suggesting that in spite of the depth of sampling, the library has not been exhausted and it is too early to predict the number of genes expressed exclusively in retina. Without doubt, the retSSH library still contains a number of valuable genes and information which awaits further discovery.

4. Evaluation of the effectiveness of the chosen approaches

The UniGene and retSSH approaches used to identify genes expressed preferentially in retina, are complementary and both have advantages and disadvantages. To evaluate what would be the most effective method for future studies, the success rate (Fig. 57) and possible setbacks of each approach needs to be compared.

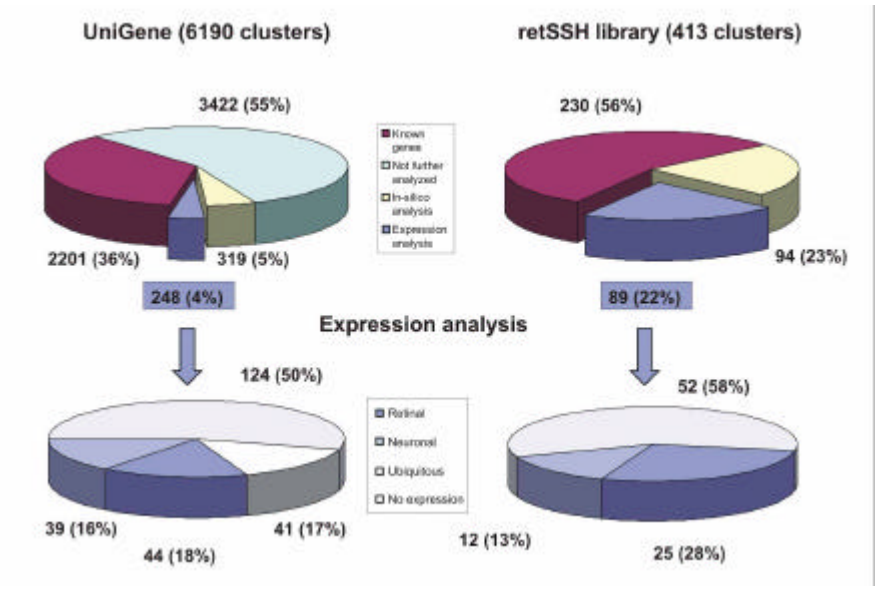


Fig. 57 Summary of results obtained with the UniGene and retSSH approaches
As of June 2000, 17,736 retina ESTs were grouped in 6190 clusters in the UniGene database. After exclusion of 2201 clusters representing known genes, the EST content of the remaining 3989 clusters was computed. A total of 567 clusters whose retina EST content was greater than 30% were investigated in detail using biocomputing tools and 248 were selected for expression analysis. Using RT-PCR retinal expression was confirmed for 44 of the clusters, whereas 39 had neuronal expression.
A total of 413 unique transcripts were identified from the retSSH after sequencing 1080 clones. Expression profiling of 89 of these clusters with RT-PCR led to the identification of 25 retinal and 12 neuronal genes.

The retSSH identifies the highest proportion of genes preferentially expressed in retina. From the 89 expression profiles done for genes identified from the library, 41% were expressed exclusively in retina or in the nervous system versus 34% of the transcripts tested in the UniGene approach (Fig. 57). It is interesting to note that the proportion of retina-specific genes is much higher in the retSSH approach (28% vs. 18%), but the number of neuronal genes is lower (13%, vs. 16%). The percentage of ubiquitous genes is also higher for the library approach, but this is only due to the fact that all of the amplifications of the retSSH project amplified a product from at least one tissue. In contrast, 17% of the PCRs of the UniGene project amplified a product from the genomic control but none from the cDNA templates.

This latter issue is one of the disadvantages of the UniGene approach and is probably due to a high content of artefactual sequences in the publicly available databases. Therefore, many of the clusters containing just one EST or more ESTs but all derived from only a single library may not represent real transcripts but may result from clones containing contaminating genomic DNA or unspliced mRNAs. The advantage of the retSSH approach is the enrichment for rare transcripts which are normally underrepresented in the public EST databases (Romualdi et al. 2001). Therefore, based on the effort needed to isolate novel tissue-specific genes and the outcome obtained with both approaches, the use of a SSH library is recommendable for such projects.

5. Expression profiling of genes

Within the scope of the Human Genome Project and countless structural genomic enterprises, it has been possible to identify thousands of genes in the past five years. However, detailed information about gene structure is not sufficient to embark on the next goal of functional studies. A pre-requisite for proper functional genomics will be to expand our limited knowledge of human gene expression. Nowadays, various techniques are routinely used for expression profiling. These include RT-PCR, classic and virtual Northern (VN) blot analyses, microarray analysis, and qRT-PCR. Indirect, estimative approaches such as SAGE and Digital Differential Display (DDD)[45] are also frequently used. Although all techniques are valuable each has certain limitations and biases. Therefore, to identify genes expressed specifically or preferentially in the retina a combination of RT-PCR, VN blot analysis, and qRT-PCR was used. This allowed us to determine the expression profile of the genes with a good level of confidence and provided experimental information to evaluate each method.

5.1. Expression profiling using RT-PCR

Because of its flexibility, RT-PCR was used as the screening technology. The advantage of RT-PCR is a minimal requirement to optimize the reactions. In most cases good semi-quantitative results can be obtained. The amplification of 337 genes from a panel containing cDNA from different tissues led to the identification of 69 retina-specific genes, 51 neuronal genes, and 176 ubiquitously expressed genes. The remaining 41 primer pairs amplified a genomic product but none of the cDNAs.

[45] http://www.ncbi.nlm.nih.gov/UniGene/info_ddd.shtml

A frequent observation from the results of the expression analyses was that a high proportion of genes expressed in retina or brain were also expressed in testis. A biological explanation for the expression of so many genes in testis might be the need to exchange histones against the more tightly packing protamines in spermatogenesis (P Burgoyne, cited in Wilda et al. 2000). The authors also propose that the expression of a set of tissue-specific genes is vital to achieve speciation and postulate that brain, testis, and placenta genes play a key role.

5.2. Expression profiling using VN blot

The expression of all retina-specific or neuronal genes identified from the retSSH project was validated by VN blot analysis (Franz et al. 1999). This alternative to conventional Northern blotting has the advantage that only minute amounts of RNA are needed since the hybridization is to immobilized full-length cDNAs. It requires more hands-on time as RT-PCR since it is necessary to perform many optimizations. On the other hand, it has the advantage that it is possible to determine not only the expression profile but also to analyze the size of the transcript. This is especially important for genes presenting various isoforms, some of which may have specific expression profiles. Using RT-PCR, if a fragment common to some or all of the isoforms is PCR-amplified, the expression of the specific isoform is overlooked.

The reliability of VN blot analysis in comparison to Northern blot analysis has been investigated. Both methods produce the same results as long as the transcripts are not too long (Hämmerle et al. 2003). This has been observed by Endege et al. (1999) who report that the SMART PCR cDNA procedure, which is commonly used in VN blot analysis, does not generate full-length cDNA, especially for long transcripts. To date the maximal size for which consistent reproduction of total cDNA population has been seen is 4 kb (Oduol et al. 2000). This phenomenon might be due to the limitations of the polymerases used, which have a theoretical copying length limit of 6 kb, but it is clear that probably very few copies of such long mRNAs are produced. From the data obtained in our analysis it appears that this method can be successfully used for transcripts up to 2.5 kb. This may explain why it was not possible to determine the expression of 14 genes. Seven of them have a reported transcript size greater than 2.6 kb and the full-length sequence of the remaining genes is not known and may eventually also exceed 2.5 kb.

An analysis of the correlation between the RT-PCR and the VN blot hybridization results obtained in this project shows that both methods produced the same results for the retina-specific genes (Table 34). Since the VN blots did not include brain cDNA, it was not possible to confirm expression in brain. Nevertheless, all genes which had a neuronal expression profile with RT-PCR were expressed only in retina in the VN and therefore the neuronal expression would have been probably confirmed if brain had been included. The results obtained by VN analysis allowed us to verify the reported size of seven transcripts and determine the transcript sizes for 12 genes (Table 19). For three genes (L18, L21, and L86) with a reported transcript size longer than 2.5 kb, a discrepancy between the reported and the

observed transcript size was found. As mentioned above, this may be due to difficulties with the amplification of longer transcripts.

5.3. Expression profiling using qRT-PCR

This technology, developed in the 1990s (Higuchi et al. 1993, Bassler et al. 1995, Heid et al. 1996) allows product amount monitoring at each cycle of the PCR amplification. It is accurate, precise, and relatively easy to perform. It not only allows the detection and quantification of rare transcripts but also the quantification of small changes in gene expression. The data obtained by this method are only reliable, if proper optimization and controls are carried out. Interpretation of results with little understanding of the intricacies involved in the generation of the data can easily lead to misinterpretation of the results. Therefore it is recommendable to follow the guidelines suggested by Bustin (2002) which include: a. Strict validation of reagents, b. Generation of a standard curve for every qRT-PCR, and c. Normalization of results against a panel of housekeeping genes whose expression is unaffected by experimental conditions.

All real-time PCR systems rely upon the detection and quantification of a fluorescent reporter, the signal of which is in direct proportion to the amount of PCR product in a reaction. A number of detection chemistries have been developed, with novel ones being constantly introduced to the market. These include intercalating dyes (SYBR Green), hydrolysis probes (TaqMan), hybridisation probes (FRET), and molecular beacons. The simplest and most economical intercalating dye format uses SYBR Green as a reporter. This minor groove binding dye binds only to double-strand DNA. It is highly sensitive since upon binding the emitted fluorescence increases over a hundred fold. A major disadvantage is that it will bind to any double-stranded DNA in the reaction, including primer-dimers and other non-specific reaction products, which may result in an overestimation of the target concentration. Nonetheless, for PCR reactions with well designed primers and optimized PCR conditions, SYBR Green can work extremely well, and therefore was the detection method chosen for this project.

Real-time quantitative PCR can be applied to function either as an absolute or relative quantification method. In order to perform relative gene expression comparisons it is necessary to determine the expression of an endogenous control, whose expression should remain constant between the different samples. Logically, housekeeping genes are the best candidates to fulfill such a condition, but there is still a lot of debate in respect to which and how many genes should be exactly used (Bustin 2000, Suzuki et al. 2000, Vandesompele et al. 2002, Tricarico et al. 2002). The literature reports that housekeeping gene expression can vary considerably between different samples (Thellin et al. 1999). Therefore for accurate normalization it is vital to have a clear understanding of the practical problems, do careful experimental design, application and validation.

In spite of the warnings, a comprehensive literature analysis of expression studies that were published in high-impact journals during 1999 indicated that G3PDH, ACTB, 18S and 28S rRNA were used as

single control genes for normalization in more than 90% of cases (Suzuki et al. 2000). Normalization to just one gene overlooks the fact that all genes are expressed at varying levels in different cells (Fig. 58) and therefore the normalization using only one gene undoubtedly leads to misleading results. Given the large tissue heterogeneity of our panel, we took special care to select the best control genes following the recommendations by Vandesompele et al. (2002). From a set of genes with different transcript abundance (GUS, ACTB, B2M, HMBS, HPRT1, RPL13A, SDHA, TBP, YWHAZ and G3PDH) we chose all except YWHAZ, G3PDH, HPRT1, and HMBS for normalization of our cDNA panels.

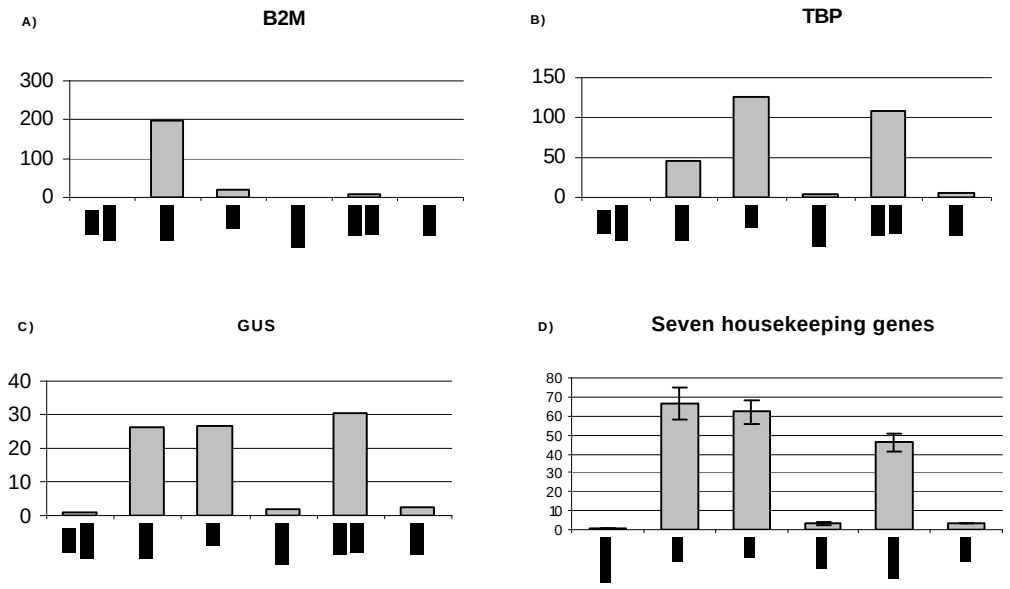


Fig. 58 Results obtained by normalization with different genes
The results of a gene expression analysis were normalized using various housekeeping genes as reference. Whereas normalization using only B2M (A) would result in the categorization of the gene as retina-specific, the normalization with either TBP (B) or GUS (C) would lead the researcher to state that the gene is expressed in retina and RPE as well as in distal colon. The same results normalized using the normalization factors calculated using seven genes (D), determined that the gene is expressed in retina, RPE, and distal colon. Data for the figure was collected and analyzed by Christine Wiedemann.

Another key aspect of qRT-PCR is the issue of reaction efficiency. The mainstream tendency is to assume that the efficiency of a reaction is always optimal. In practical terms this means that in the exponential phase the amount of product will be doubled in each cycle. Actually this is not the case for most PCRs because numerous factors such as PCR inhibitors or secondary structure decrease the efficiency. Since the amplification efficiencies of the reference and target genes are seldom similar, the use of the broadly adopted second derivative ($2^{-??Ct}$) method could lead to substantial errors. The average efficiency of our reactions was 1.86 (93%), but values ranged between 1.58 and 2.08 (Table 37). The latter value points out to amplification of primer dimers especially from the samples with low template concentration. To calculate the relative amount of transcript we therefore used a refined version of the second derivative formula that takes into account the amplification efficiency of target and internal control genes.

As already mentioned, primer-dimers are the greatest drawback of using the SYBR Green detection method. But the combined use of several strategies can help control as much as possible this factor. This includes proper primer design, evaluation of the most stringent and efficient buffer composition, setup of the reaction using hot-start conditions, and selection of proper annealing and fluorescence-reading temperatures. The optimization of these factors is the main reason why qRT-PCR is more work-intensive than conventional RT-PCR. Therefore, from our experience it is not recommendable to start expression profiling directly with qRT-PCR because without knowledge about the tissues in which a gene is expressed the optimization is very difficult and time-consuming.

In the final phase of expression profiling, we applied qRT-PCR to quantitate the mRNA levels of 52 genes identified in the course of the project as retina-specific or neuronal. The expression of these genes was investigated in panels that included cDNAs from tissues of all embryological origins. Since a study of mouse full-length transcripts (Carninci et al. 2003) found that there is greater diversity between the central nervous system (CNS) and retina, inner ear, and other peripheral nervous tissue as previously thought, we decided to refine the neuronal expression analysis by including samples from anatomically distinct regions of the CNS (basal ganglia, cerebellum, and occipital cortex). The inclusion of occipital cortex was dictated by the fact that this is the brain region where the visual information is processed and we wondered if there are genes that are expressed exclusively in retina and occipital cortex.

Primers for 58 genes were designed but the expression of six of these genes (A038, A170, A180, A199, B043, and L86) could not be determined because of primer-dimers, poor reaction efficiencies, or presence of unspecific products. The expression profiling of the remaining genes determined that 18 are retina-specific, 20 neuronal, seven ubiquitous, one is retina- and heart-specific, one is retina- and stomach- specific, and four are expressed primordially in the RPE (Table 34). Twelve of the 20 neuronal genes had higher expression in at least one of CNS regions than in retina or RPE. For most genes the expression levels in the various central nervous system regions were not very different, but there were some exceptions. For example, A111 (BC016878) was categorized as a retina-specific gene by RT-PCR, but the examination with qRT-PCR determined that it is expressed at high levels in retina, RPE, and occipital cortex. In contrast, the concentration of this transcript in basal ganglia and cerebellum is not higher than in distal colon, lung, or stomach. High expression in cerebellum and retina, with very low expression in basal ganglia, occipital cortex, and RPE was found for only one gene, A109 (KIAA1263). A survey of the expression levels of the genes in the three CNS samples, determined that the highest expression is usually observed in cerebellum and occipital cortex.

The majority of genes preferentially expressed in retina or neuronal tissues are not expressed in RPE. But we did discover four genes which are expressed at high levels in RPE (Table 34). Three of them (L21, L78, and L39) have higher expression in RPE than any other tissue. For a fourth gene (L93) the expression level in RPE is only slightly lower than in retina.

Table 34 Summary of expression profiles obtained with various methods

Lab ID	Gene ID[a]	RT-PCR	VN[b]	qRT-PCR[b]	Lab ID	Gene ID[a]	RT-PCR	VN[b]	qRT-PCR[b]
A004	CAMTA1	N	-	N	L18	FLJ13305	AR	R	R
A017	ERO1LB	R	-	U	L20	PHAX	AR	AR	-
A059	PSMD11	R	-	U	L21	FLJ23460	N	N?	RPE+O
A084	KIAA1796	N	-	N	L23	C20orf103	N	N?	N
A085	LOC157627	N	-	N	L24	SLC1A2	N	N?	N
A106	LMAN1L	N	-	R	L25	ABCC5 variant	U	R	-
A109	KIAA1263	N	-	N	L27	HSPC159	AR	R	-
A111	BC016878	R	-	N	L28	SV2B	AR	none	R
A126	C12orf3	R	-	R	L30	RIPX	AR	none	-
A150	DKFZp547H074	N	-	N	L32	FLJ31121	U	none	-
A165	CHD1	R	-	U	L33	uL33	R	R	R
A166	C1orf32	R	-	N	L35	uL35	R	none	R
A168	ORC2L	N	-	U	L36	uL36	AR	none	R
A169	AA057097	R	-	R	L37	uL37	R	R	R
A177	DKFZp761D221	N	-	N	L38	uL38	R	R	R
A203	STK35c	R	-	R	L39	C4orf11	R	R	RPE+R
A205	BC035234	R	-	R	L40	ZPBP	R	R	N
A206	AK056484	N	-	N	L47	MGC14816	R+H	R+H	R+H
A211	SF3B3a	R	-	U	L48	uL48	R	R	R
A213	CRYPTIC	N	-	R+S	L50	uL50	R	R	-
B001	FLJ31564	N	-	N	L52	BC040189	N	-	R
B015	C14orf29	N	-	N	L54	uL54	AR	R	R
B030	KIAA1917	N	-	N	L56	uL56	R	R	R
L02	CLASP2	AR	none	N	L63	DKFZp547C176	N	none	N
L03	SOCS5	U	none	-	L72	H2AV	N	none	N
L05	DKFZp761F0118	AR	none	U	L78	FLJ30499	N	none	RPE+R
L10	C14orf129	AR	AR	AR	L86	KIAA1013	R	AR	-
L11	KIAA1576	N	-	N	L88	KIAA1579	R	none	U
L14	FLJ33282	N	none	R	L92	DKFZp547J1816	N	none	-
L16	PLCD4	R	AR	R	L93	DAPL1	N	N?	R+RPE
L17	SSX2IP	AR	N?	N					

[a] The gene ID indicates the sequence from which the primers for the qRT-PCR or the VN were designed. A gene ID beginning with uL indicates that the primers were designed based on the sequence of a clone which has not been made public yet.
[b] AR: abundant in retina; H: heart; N: neuronal; N?: expression is probably neuronal; O: other tissues; R: retinal; RPE: retinal pigment epithelium; S: stomach; U: ubiquitous; none: no hybridization signal; - : was not done.
[c] The sequences probably correspond to novel splice variants of the gene
Expression profiling by RT-PCR of the genes with a lab ID starting with 'L' was done by Jelena Stojic. Results were obtained by personal communication.

It was surprising to find seven ubiquitously expressed genes in the set of neuronal and retina-specific genes (Table 34). For five of the genes this discrepancy can probably be explained by careful consideration of the fragment of the gene amplified with each method. In the case of A017, A059, A165, A168, and A211 when the original amplification was done, no gene organization was known and the primers were designed based on the EST sequence. By the time the primers for qRT-PCR were designed, the gene structures were known and we found that the original primers align to sequences mapping to the introns of the probable gene. Therefore, the products amplified by each method differ since the qRT-PCR primers were designed to amplify exonic sequence of the corresponding gene. For example, the sequence from A017 amplified by RT-PCR extends from the sixth exon into IVS6 whereas the qRT-PCR primers amplify exons 15 and 16 of ERO1LB. Thus, the seemingly confounding expression profiles obtained for these genes are probably due to the analysis of different isoforms. To investigate if tissue-specific isoforms of these genes may exist it would be recommendable to design new primers and analyze with qRT-PCR the original isoforms investigated by RT-PCR. The expression of L05 in retina, measured by RT-PCR, was not so much higher than in the other tissues, and therefore the qRT-PCR expression profile is probably reliable. Therefore, actually only one discrepancy (for L88) was seen after investigating the expression of 61 genes using

two or three different methods (Table 34). This confirms that for general and quick evaluation purposes, RT-PCR is sufficient.

5.4. Summary of the expression profiling effort

As a result of the comprehensive expression profiling achieved in this project, we were able to identify 60 retina-specific genes, three genes expressed primordially in the RPE, and 48 genes expressed in retina and the CNS. As already discussed, there may be some more tissue-specific variants which account for higher numbers if only the RT-PCR results are considered, but they have not been further investigated. Since the expression of the 111 genes was not previously known, the collection of these candidate genes constitutes a valuable asset for the investigation of retinal pathologies. Future investigations could complement the present expression profiles with *in-situ* hybridization and immunohistochemistry analysis in order to establish the temporal and spatial expression profiles of the genes.

6. Correlation of gene structure and expression

Transcription is a slow and energy-consumptive cellular process, requiring approximately one second to transcribe 20 nucleotides at the expense of at least two ATP molecules per nucleotide (Lehninger et al. 1982). As already mentioned, pre-mRNA contains not only exonic but also intronic sequence. Intron sizes vary considerably within a genome as well as between different species with human genes having the most and largest introns (Deutsch and Long 1999 and Lander et al. 2001). In the last decade the study of intronic sequence has shown that these sequences are more relevant than previously thought and that their study may lead to interesting findings.

Recently a correlation between intron size and gene expression has been reported (Castillo-Davis et al. 2002, Versteeg et al. 2003, Eisenberg and Levanon 2003). The study by Castillo-Davis et al. found that introns in highly expressed genes are 14 times shorter, on average, than those in genes that are expressed at low levels. They report that abundant genes have introns of 343 bp versus the 4807 bp introns found in rare transcripts. Moderately expressed genes (represented by 200-2000 ESTs) have average introns of 2153 bp. There is also a decrease in exon length in highly expressed genes, but the difference is not as significant. The intron length correlation was also found by Versteeg et al. (2003). They report that not only the transcript abundance directly correlates with intron length, but the specificity of a gene is also related to intron size. A comparison of intron length of 532 housekeeping and 5404 non-housekeeping genes found that the average intron length of housekeeping genes is 2573 bp whereas genes with restricted expression have average introns of 5025 bp (Eisenberg and Levanon 2003). Interestingly, the proportion of intronless genes among genes expressed at high and low levels is similar in ubiquitous and rare transcripts (Castillo-Davis et al. 2002). Thus, it seems that short introns, but not the absence or loss of introns, are favored in highly expressed genes.

To evaluate if the principle of tissue-specificity and intron length also applies to the genes characterized in this project, the intron and exon length of six genes (C7orf9, C12orf7, C1orf32, C14orf29, C4orf11, and DPAL1) was computed. As can be observed in Fig. 59, the average exon lengths of the genes range from 132 bp (C14orf29) to 502 bp (C1orf32). The average intron length from the six genes (33 introns) is 3548 bp. This value is between the 2573 and 5025 bp lengths found for housekeeping and restricted expression genes, respectively (Eisenberg and Levanon 2003). DAPL1 and C1orf32 contain the longest introns (Fig. 59). The intron length of C12orf7 is surprisingly short with no clear correlation to its retina-specific expression. In this case the short intron length is perhaps determined by a high expression and not its specific expression. In the future it might be worthwhile to compute the length of introns and exons of all retina-specific and neuronal genes identified to date to see if this information could be used as selection criteria for future expression analysis screenings.

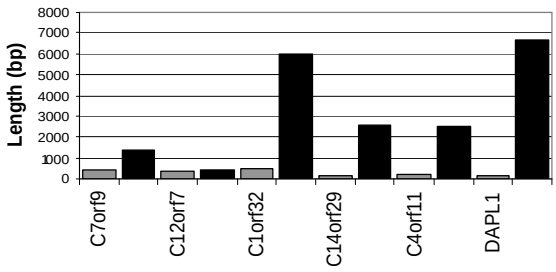


Fig. 59 Average exon and intron lengths of six genes
Average exon (grey columns) and intron (black columns) length of six genes cloned in the present project.

7. Cloning and characterization of genes expressed preferentially in the human retina

The central aim of this project was the partial characterization of the human retinal transcriptome with a special emphasis on those genes that are preferentially expressed in this tissue. Because of the high cellular heterogeneity of the retina, its transcriptome is probably very complex and as demonstrated by our analysis of various libraries, the depth of sampling needed to characterize it is still far from being reached.

Our gene identification project was started at a time when only 16% of the genome had been finished, 47% was still available only as draft sequence, and 37% had not been sequenced yet. Since then, gene identification and characterization have become more feasible due to the completion of the human genome project and the development of algorithms to identify coding regions from genome sequence. To achieve this, two main approaches are used. The first (alignment-based algorithms) relies on alignment of biological sequences to the genomic sequence. The second (*ab initio* predictions) relies on the identification of patterns in genomic DNA (e.g. introns, coding regions, splice sites, GC-rich segments, promoters, etc.) which are characteristic of genes (Burge and Karlin 1998, Stormo 2000). However, the accuracy with which genes can be predicted is still far from satisfactory (Zhang 2002). Therefore, the cloning of a gene is still a challenging issue which requires experimental work.

Gene cloning strategies used in this project included EST assembly, cDNA library screening, rapid amplification of cDNA ends (RACE), PCR amplification of exon predictions, and bioinformatical analyses. Although theoretically gene cloning might appear an easy, straight-forward enterprise, there are serveral factors which render gene characterization challenging. These factors include the existence of artefact sequences which constitute false leads, absence of specific clones in the cDNA libraries, sequence complexities (e.g. secondary structure and high GC content), base changes (SNPs), and alternative splicing. In our project, alternative splicing and the existence of various isoforms implied a particular challenge for the cloning efforts. Therefore, an overview of alternative splicing and its special significance for the retinal transcriptome will be discussed briefly before presenting and discussing particular genes.

7.1. Significance of alternative splicing in the retina

As estimates of the number of genes in the human genome fell from approximately 100,000 to 30,000, interest in alternative splicing as a source of increased diversity has soared. At the same time, the estimates of genes presenting alternative splicing have kept on climbing from 5% in 1994 to 60% in 2002 (Sharp 1994 and Modrek and Lee 2002). Even though already high, the percentage of alternatively spliced genes will probably increase even more as the depth of sampling increases and novel transcripts reveal until then unknown splice variants. A study of chromosome 21 and 22 transcripts revealed that the number of transcripts probably greatly exceeds the prediction of 30,000 expressed genes (Kapranov et al. 2002). Therefore, the simple aphorism ‘one gene - one protein - one disease’ has been long abandoned and the role of alternative splicing is becoming increasingly more important. Its significance extends beyond the ability to create protein diversity to a central role in the regulation of transcript levels (Ladd and Cooper 2002). Splicing alterations are also an important cause of disease as proven in a study of more than 32,000 mutations associated with human inherited diseases. It was observed that 14% of the point mutations were located at the intron-exon junctions and probably result in RNA splicing defects (Krawczak et al. 2000 and Cartegni et al. 2002).

RNA splicing is the essential, precisely regulated process that occurs in the nucleus after gene transcription and before mRNA translation. The precision and complexity of intron removal during pre-mRNA splicing still amazes 26 years after the discovery that the coding information is interrupted by introns (Berget et al. 1977 and Chow et al. 1977). The incredible level of genetic diversity arising from alternative splicing is best illustrated by the Drosophila axon guidance receptor, *Dscam*, which can produce as many as 38,000 mRNA splice variants (Schmucker et al. 2000). Alternative splicing may result from the inclusion or exclusion of a ‘cryptic’ exon as well as from the use of alternative splicing donor or acceptor sites. In mouse 70% of alternative splice forms are due to the use of cryptic exons and not to alternative splice sites. Fifty-six percent of these mouse cryptic exons are internal exons (Zavolan et al. 2003). Another interesting aspect is that most splice variations affect the coding region (Modrek et al. 2001 and Zavolan et al. 2002) but tend to insert or delete complete protein domains more frequently than expected by chance (Kriventseva et al. 2003). Alternative splicing might also play a role in specification. Comparative genomic studies have revealed that whereas only 2.5% of human

constitutive exons do not have a matching exon in the orthologous gene, 73% of human cryptic exons are not conserved in mouse (Modrek and Lee 2003).

Cryptic exons have been reported to be shorter than constitutive exons (Berget 1995) and to be flanked by weaker splice sites (Stamm et al. 2000). An analysis of the data from the genes cloned in this project confirms this trend. Whereas 48 constitutive exons have a mean length of 311 bp, the 14 cryptic exons are, on average, 99 bp long. The cryptic exons also contain slightly weaker splice scores with 5.62 (cryptic) versus 5.37 (constitutive) for the acceptor and 3.45 (cryptic) versus 3.30 (constitutive) for the donor sites.

Components involved in the fine intron-removing mechanism include highly conserved donor and acceptor consensus sequences, splicing enhancers and inhibitors, and the spliceosome complex. But it is clear that many more components are involved and still have to be discovered. Therefore, presently it is very difficult to predict and foresee the particular splicing pattern of a gene and novel splice variants are only discovered by experimental analyses. Additionally, a large fraction of alternative splicing undergoes cell-specific regulation but very little is known about tissue-specific alternative splicing and its regulation. In a study by Xu et al. (2002) involving the analysis of 2.2 million human ESTs, 10-30% of human alternatively spliced genes presented evidence of tissue-specific splice forms. They found that the number of tissue-specific alternative splice forms is highest in brain, which confirms a report by Stamm et al. (2000). Interestingly, if the number of splice variants is normalized to the total number of ESTs studied for the particular tissue, skin and eye_retina have the greatest enrichment of tissue-specific splicing. Both present 2.4 times more tissue-specific splice forms than average (Fig. 60), which is double than the ratio obtained for brain.

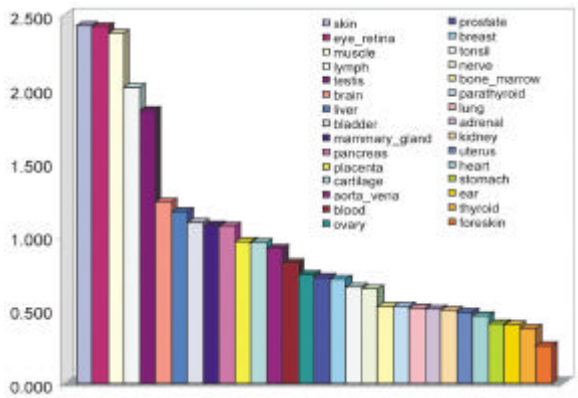


Fig. 60 Enrichment of tissue-specific alternative splicing in 30 human tissues
The *y*-axis shows the enrichment factor for each tissue, calculated by dividing the number of tissue-specific alternative splices observed in a tissue by the total number of ESTs observed in that tissue, normalized to have an average value of 1. (Reproduced from Xu et al. 2002 with permission of Oxford University Press)

The enrichment of retina-specific alternative splice forms implicates that the establishment of the retinal transcriptome will require special care and attention in order to fathom all transcripts. The high alternative splice rate of the retina was reflected in our EST analysis, which identified a number of ESTs that align either partially to exons and introns of already known genes or completely in the introns. The only explanation for this observation is that: a. the sequences resulted from genomic contamination or incomplete splicing, or b: that they constitute novel tissue-specific splice variants.

In light of these facts, efforts to elucidate the regulation mechanism of alternative splicing, with a particular emphasis on the understanding of tissue-specific splicing regulation, should be started. Since the retina presents such a high number of splicing variants, perhaps the bioinformatical analysis of retina isoforms and the corresponding introns might be valuable for the identification of regulatory motifs.

The existence of various gene isoforms has the added problem that it is sometimes difficult to determine the sequence of the encoded protein(s). Actually, it is probable that many of the isoforms, rather than be translated, may undergo regulated unproductive splicing and translation (RUST) (Lewis et al. 2003). Lewis et al. found that a third of 5693 reliably-inferred alternative mRNA isoforms are candidates for nonsense-mediated mRNA decay (NMD), an mRNA surveillance system. It is probable that the coupling of alternative splicing and NMD is used by the cell machinery to regulate protein expression in a developmental stage- and cell-specific manner. The advantage of regulation via alternative isoforms is that this post-transcriptional regulation can provide temporal control unattainable by transcription factors (Lewis et al. 2003). Since phototransduction is a very fast process requiring constant re-accommodation, it is feasible that a fast regulation of this type is vital for the normal functioning of the visual process.

From the seven genes cloned during this thesis, four (C12orf7, GRM7, C14orf29, and C4orf11) have been confirmed to contain alternative splice forms. The investigative work done for GRM7 consisted precisely in the characterization of novel splice variants with tissue-restricted expression. The cloning of C12orf7 was particularly difficult due to its intricate splice pattern.

8. Characterization of C7orf9

The cloning and characterization of C7orf9 was begun after finding that cluster Hs.60473 is specifically expressed in the retina (Fig. 27). Sequence assembly of two clones isolated from a retina cDNA library yielded a 1190 bp transcript which encodes a 196 aa protein. Shortly afterwards, the cloning of this gene was also reported by two other groups (Hinuma et al. 2000 and Liu et al. 2001) who proposed that the encoded protein contains two (Liu et al. 2001) or three (Hinuma et al. 2000) RFamide-related peptides (RFRPs).

Biologically active peptides containing an RFamide structure at their C-termini are commonly found in the animal kingdom (Yoshida et al. 2003). The RFRPs constitute a large family of neuropeptides known to exert a variety of functions such as neurotransmission, neuromodulation, cardioexcitation or control of muscle contraction (Raffa et al. 1988 and Greenberg et al. 1992. Hinuma et al. (2000) were able to demonstrate that synthetic RFRP-1 and RFRP-3 peptides are specific agonists of an orphan G protein-coupled receptor OT7T022 and therefore possibly regulate prolactin secretion. Furthermore, endogenous RFRP-1 was purified from bovine hypothalamus and found to consist of 35 aa residues suggesting that the mature neuropeptides are generated by cleavage of the RFRP preprotein

(Fukusumi et al. 2001). The concentration of the 23 aa rat RFRP-3 has also been reported to be high in the hypothalamus but surprisingly no expression was detected in the eye ball sample (Yoshida et al. 2003).

The identification of the C7orf9 gene was particularly interesting because it maps within the critical region for dominant cystoid macular dystrophy (CYMD) (Kremer et al. 1994 and personal communication) and was therefore an excellent candidate for the CYMD gene. To investigate its possible role, the complete coding region of the gene was sequenced from a control person and two CYMD patients. Although three base changes (c.96G>C, c.125A>G, and c.384C>T) were indeed found in the patients, none of them are related with the pathology since the minor alleles are also found in healthy individuals at a frequency of 7 to 27%. A possible heterozygous deletion of the entire gene in the patients could also be discarded because both patients were heterozygous for one of the mentioned single nucleotide polymorphisms. Since gross intragenic rearrangements were also absent in the patients, we conclude that the C7orf9 gene is possibly not involved in the etiology of CYMD. It should be noted, however, that mutations in the promoter sequence or within the intronic regions which could lead to changes in the level of expression or splicing have not been investigated.

9. Characterization of C12orf7

The cloning and characterization of the retina-specific C12orf7 gene was achieved by using several different approaches such as cDNA library screening, 5'-RACE, and PCR amplification. Six exons ranging in length from 94 to 536 bp and presenting 13 different splice variants could be identified. In addition, we found that IVS4 and IVS5 are retained in some transcripts. Particularly the inclusion of IVS5 seems to be favored since by PCR amplification a higher fraction of intron-retaining product was observed.

In spite of the repeated efforts to clone the 5'-end of the gene, it was not possible to prolong the present sequence so that it would include an in-frame stop codon. Comparative genomic analysis have revealed that an alternative initial exon may exist since there are two regions with high identity between the human, mouse, and rat genomic sequences in a region 2.8 kb upstream of the current exon 1. On the other hand, the low conservation of exon 1 between the species would indicate that all the exons are already cloned.

Because none of the isolated clones contained a full-length sequence and PCR amplification of the complete transcript was not possible, it is difficult to estimate the exact nature of the *in vivo* transcripts. Nevertheless, there is solid evidence of at least 11 different isoforms which differ in their content of exons 1 thru 5. Therefore, by combination of the 11 different splice forms to the two possible 3'-ends it is feasible that at least 22 isoforms of this gene are expressed in the cell. From the assembled isoforms the one with the longest cDNA, C12orf7_v1, also encodes the longest ORF (471 aa). Additional evidence that the splice variants are real and have biological significance comes from the fact that the majority of them are also conserved in mouse.

The longest protein encoded by the C12orf7 gene is characterized by the presence of a nuclear localization signal and five ankyrin repeats, some of which are lost in the alternative splice variants. Ankyrin repeats are tandemly repeated modules of about 33 amino acids which are found in numerous proteins mainly from eukaryotes. In humans more than 325 unrelated proteins contain this repeat (Mohler et al. 2002). The proteins containing ankyrin repeats are found in the nucleus, cytoplasm, and the extracellular space and have diverse functions, such as transcription initiation, cell cycle regulation, cytoskeletal integrity, ion transport, and cell-cell signalling (Sedgwick and Smerdon 1999). To date, ten high-resolution structures of ankyrin repeat proteins have been solved. They closely resemble one another despite their different cellular functions, supporting the role of ankyrin repeats in protein-protein interactions (Mosavi et al. 2002). The biophysical and X-ray crystallographic studies of proteins containing three or four repeats demonstrated that they are well-folded, monomeric, posses high thermostability, and adopt a very regular, tightly packed ankyrin repeat fold (Mosavi et al. 2002). The ankyrin repeat consists of antiparallel alpha helices stacked side by side which are connected by a series of intervening beta hairpin motifs. This core structure is sufficiently robust to withstand a considerable degree of sequence variation at the amino acid level while maintaining a stable framework for presenting surface contact residues to mediate the protein-protein interactions (Sedgwick and Smerdon, 1999). Essentially, all well-conserved positions are located in the interior of the ankyrin repeat and therefore their primary role is structure formation rather than binding specificity (Mosavi et al. 2002). The ability of ankyrin repeats to bind target proteins commonly involves contacts formed through the tips of the beta hairpins and the surface of the helical bundle facing the ankyrin groove.

At least two genes expressed in retina which encode proteins containing ankyrin repeats have already been identified. The first is the retina-specific rdgA gene, which was identified from the study of the Drosophila retinal degeneration mutant A (Masai et al. 1993). This model animal has photoreceptor cells that differentiate normally but degenerate rapidly after eclosion (Kanoh et al. 1983). Biochemical studies demonstrated that the gene encodes a membrane-associated diacylglycerol kinase which is anchored to a membrane, possibly by the ankyrin repeats. Studies of a severe degeneration allele suggested that the ankyrin repeats are necessary for normal functioning of the DGK protein (Masai et al. 1993).

The second gene, published just recently, is the gene encoding Sans (Kikkawa et al. 2003). Mutations in this gene, which is expressed in cochlea, brain, cerebellum, eye, and testis, have been found to cause Usher syndrome type I G (Weil et al. 2003). This syndrome affects both the inner ear and the retina.

It is significant that the shorter isoforms of C12orf7 encode ORFs which are partially similar to the longest ORF but many contain a premature termination codon (PTC) and are therefore candidates for NMD. The widespread coupling of alternative splicing and NMD in humans has already been demonstrated (Lewis et al. 2003). NMD results from the activation of a surveillance mechanism that is ubiquitous among eukaryotes and which leads to accelerated mRNA decay, particularly of transcripts

containing a PTC. Although the role of NMD was originally thought to be the monitoring of transcription errors, it is now widely accepted that this mechanism is also involved in the control of gene expression by regulating the stability of selected physiological transcripts (Dahlseid et al. 1998, Culbertson 1999, Frischmeyer and Diezt 1999). Therefore, some of the direct targets of the NMD pathway probably function as transcriptional regulators. Concrete examples backing this hypothesis include the regulation of expression via NMD for glutaminase (Labow et al. 2001), fibroblast growth factor receptor 2 (Jones et al. 2001), and ABCC4 (Lamba et al. 2003). The high variety of C12orf7 isoforms coupled to the high conservation of the alternative exons in human, mouse, and rat surely demands in-depth investigation in order to determine the biological significance of each isoform and the possible involvement of NMD in the regulation of transcript level.

The retina-specific expression of C12orf7, the existence of the ankyrin repeats as well as the nuclear localization signal suggest that this gene could play a role in the structure or the maintenance of the nuclear envelope. Therefore the establishment of the subcellular localization of this protein, the identification of its interaction partners, and the investigation about the nature and role of the various isoforms should constitute the next research objectives.

10. Characterization of GRM7

In the mammalian CNS, glutamate mediates excitational neurotransmission via ionotropic and metabotropic receptors. Metabotropic glutamate receptors (GRMs, formerly mGluRs) belong to the superfamily of G protein-coupled receptors and currently comprise a family of eight distinct subtypes (GRM1 to GRM8). For six of the eight family members (GRM1, GRM4, GRM5, GRM6, GRM7, GRM8) several variants have been reported, most of which are generated by alternative exon usage leading to distinct carboxyl-terminal domains. The general structure of the metabotropic receptors consists of a glutamate binding site, a cystein-rich region, a seven transmembrane domain and an intracellular C-terminal region.

GRM7 was originally cloned from a rat forebrain cDNA library (Okamoto et al. 1994). Subsequently, the human homologue was identified (Makoff et al. 1996) and later two isoforms, GRM7_v1 and GRM7_v2 (formerly mGlu7a and mGlu7b) were reported (Flor et al. 1997). The two transcript variants differ by an out-of-frame insertion of 92 bp at the 3'-end of the coding region resulting in two putative proteins of 915 and 922 aa with distinct C-termini. The cloning of five additional human variants, three of which encode proteins with novel C-termini was achieved within the scope of this doctoral project. The GRM7_v3 isoform encodes the longest putative protein (924 aa) reported until now for this gene. The putative GRM7_v4 and _v5 proteins are shorter, with 911 and 906 aa, respectively.

Glutamate receptors are principally found in neuronal tissues but as demonstrated by our expression analyses the expression of GRM7 is not limited to the CNS (Fig. 36). In the neuronal tissues all five isoforms are expressed, but the isoform abundance in each tissue varies. Whereas in brain the expression levels of all transcripts is approximately the same, in retina there is much stronger

expression of GRM7_v2, _v3, and _v4. The non-neuronal tissues with high expression of one or more variants include trachea, testis, uterus, placenta, and adrenal gland. This extra-neuronal expression can be explained by the recent finding that glutamate may also act as a signaling molecule in paracrine processes (reviewed in Skerry and Genever 2001). While the GRM7_v1 isoform is expressed throughout the CNS (Kinoshita et al. 1998), GRM7_v2 localization in the brain appears to be more restricted and is preferentially found in distinct regions such as hippocampus, ventral pallidum, and globus pallidus (Kinoshita et al. 1998). Both isoforms are localized to the active zones of presynaptic axon terminals close to the glutamate release site (Saugstad et al. 1994 and Kinoshita et al. 1998). Electron microscopy further demonstrated that GRM7_v1 is present asymmetrically at pre- and postsynaptic sites at certain cone bipolar cell ribbon synapses possibly reflecting functional activity in the differential activation of postsynaptic neurons of the inner plexiform layer (Brandstätter et al. 1996).

Although the precise physiological functions of the GRM7 subtypes are still unclear, targeted disruption of the orthologous murine Grm7 gene locus has been shown to cause a deficit in fear response and an impairment of taste aversion. This suggests a role of Grm7 in amygdala function which is essential in relating these behavioural traits (Masugi et al. 1999). Furthermore, upon drug induction mice lacking Grm7 are susceptible to epileptic seizures indicating that Grm7 may be particularly important in the regulation of neuronal excitability (Sansig et al. 2001). There is strong evidence that the GRM7 receptor acts as an autoreceptor to provide negative feedback for the release of glutamate. Supporting this hypothesis is the fact that GRM7 has a low affinity for glutamate and a preferential presynaptic localization (Schoepp 2001).

The C-terminal domain of GRM7_v1 appears to be divided into three functional regions. The most proximal encompassing amino acid residues 856 to 878 plays a role in the signalling complex (Nakajima et al. 1999), the central part (aa 883-915) is involved in the axonal targeting (Stowell and Craig 1999), while the immediate C-terminus (aa 912-915) is involved in the synaptic clustering of the receptor via PDZ domain proteins (reviewed in Dev et al. 2001).

GRM7-mediated neurotransmission depends critically on its regulation by associated molecules, such as enzymes, scaffold proteins and synaptic anchor proteins (Perroy et al. 2002 and Enz and Croci 2003). Several of the proteins interacting with the C-terminus have been identified. Those binding to both GRM7_v1 and _v2 include filamin A (Enz 2002), protein kinase C, alpha binding protein (PICK1) (Staudinger et al. 1997), and syntenin (Hirbec et al. 2002). Whereas the catalytic ?-subunits of protein phosphatase 1C (PP1C) interact selectively with GRM7_v2 (Enz 2002), glutamate receptor interacting protein (GRIP1) interacts only with GRM7_v1 (Hirbec et al. 2002). The exact binding domains of each of these proteins have been recently mapped to the last 13 aa of the rat mGluR7b protein and it was discovered that they partly overlap (Enz and Croci 2003).

The functional properties of the specific C-terminal ends of the novel GRM7 isoforms remain to be established. There is evidence that axon targeting is mediated by the distal 60 aa of the GRM7_v1

subtype (Stowell et al. 1999). Hence, the novel isoforms may be directed to distinct subcellular localizations in the respective tissues. Since GRM7-mediated signal transduction depends largely on its association with regulatory proteins, it is essential to establish the binding motifs of the novel isoforms and interacting proteins. It may be speculated however that similar to GRM7_v1 and GRM7_v2 which have been shown to contain PDZ-binding motifs at the extreme C-terminus, similar interactions may also be ascribed to the new GRM7 subtypes. In support of this notion, close inspection suggests that the C-terminal Q-S-N-L motif of GRM7_v3 may constitute a class-I PDZ motif.

Further studies need to address the physiological significance and pharmacological targets of the novel GRM7 isoforms. Based on the available scientific evidence, the GRM7 receptors represent important therapeutic targets because of their role in the regulation of normal synaptic activity. They could therefore be of great interest for therapeutic intervention (Brandstätter et al. 1996).

11. Characterization of C1orf32

The investigation of a 70 kb locus from chromosome 1q24.1 containing two retina enriched UniGene clusters, which actually derive from a single gene, led to the identification of C1orf32. The 4954 bp cDNA was assembled from clones identified from library screenings, PCR amplifications products, publicly available ESTs and 5'-RACE experiments. The use of the newly emerging field of comparative genomics was also a great aid since the sequence of seven from the ten exons was not found in any clone and was identified only as a result of the high homology between the human and mouse sequences.

Expression analysis of human C1orf32 determined that it is highly expressed in retina, fetal brain, and testis. Lower expression was seen in basal ganglia, occipital cortex, spinal cord, thymus, prostate, and uterus. A mouse Northern blot analysis confirmed that the gene is also expressed in mouse and that the transcript is over 6.5 kb in size (Fig. 40).

The C1orf32 human gene encodes a 639 aa protein which is 87% similar to the 651 aa mouse hypothetical protein. A tentative rat orthologous sequence, also assembled by comparative genomics, presents 85% similarity to the human sequence. The human and mouse proteins contain a signal peptide, an immunoglobulin domain, a short trans-membrane domain, and a cystein-rich stretch. Unfortunately, no particular functional role can be inferred based on the domains because they are all found in many different proteins. But important clues about the functional relevance of C1orf32 came from sequence comparisons which found that the human C1orf32 protein is 34% similar to the human LISCH7 protein.

LISCH7 presents an overall protein structure similar to C1orf32 characterized by an immunoglobulin domain, a transmembrane helix, and a cystein-rich region (Yen et al. 1999). It is reported to contain two di-leucine routing signals which are also present in C1orf32 but at a different position. Differences with C1orf32 include the absence of a signal peptide, the presence of an NPGY motif, and the

existence of a tumor necrosis factor, alpha (TNF-a) receptor signature. Nevertheless, an alignment of the human, murine, and rat C1orf32 and LISCH7 proteins (Fig. 42) revealed an astounding conservation between both proteins and the three species. Not only are the immunoglobulin, transmembrane and cystein-rich domains conserved, but also other stretches of the sequence are almost identical in the six sequences. They probably constitute novel motifs since searches in all motif and domain databases did not reveal similarity with already known motifs. Since the detection of conserved patterns of amino acids in related proteins often provides insight into structurally or functionally important features of the proteins, this finding constitutes an excellent starting point for future investigations.

The liver-specific bHLH-Zip transcription factor (Lisch7) gene was originally identified in rat in 1999 by Yen et al. and received this designation because it is primarily expressed in liver. In rat, the two subunits generated by alternative splicing have been found to form a lipolysis-activated receptor (Yen et al. 1999). Actually, the receptor had already been identified in 1992 by its binding capacity to low density lipoprotein in the presence of free fatty acids but it took seven years to clone the gene. Another study has also reported that the receptor binds not only to apolipoprotein B and apolipoprotein E (ApoE) but also displays great affinity for triglyceride-rich lipoproteins (Bihain and Yen 1998). The most recent finding is that along with genes involved in cell adhesion and signalling, the LISCH7 gene is a primary target of p53 regulation (Kannan et al. 2001).

The ApoE glycoprotein is found in a number of circulating plasma lipoprotein complexes and is the major apolipoprotein of the CNS. Its primary function appears to be that of a recognition ligand for receptor-specific removal of lipoproteins from circulation. In the CNS it also plays a role in the transport of cholesterol. While most ApoE is synthesized in the liver, other tissues, including brain and retina, are also capable of producing the protein. In retina, ApoE seems to be synthesized in the Müller cells (Amaratunga et al. 1996) but ApoE immunoreactivity has been found in the photoreceptor outer segments, the retinal ganglion cell layer, the retinal pigmented epithelium, and both collagenous layers of Bruch membrane (Anderson et al. 2001).

Several ApoE isoforms exist and their affinities for specific lipoprotein receptors vary. The interaction affinity of ApoE to the LISCH7 receptor also seems to be isoform-dependent since an *in vitro* experiment with very low density lipoprotein (VLDL) isolated from a Type III hyperlipidemic patient (ApoE e2/e2 phenotype) failed to bind to the receptor (Yen et al. 1994).

Large population analyses have shown that the inheritance of the e2/e2 ApoE alleles correlates with a higher risk for atherosclerosis, Alzheimer disease, and, most recently, to the incidence of age-related macular degeneration (reviewed in Stöhr 2003). Although the symptoms of the last two disorders differ, both share the fact that they are neurodegenerative diseases which manifest themselves in older persons. Additional evidence for the role of ApoE in AMD comes from a study of serum lipoprotein levels which found that AMD patients had higher ApoE concentrations than controls (Abalain et al. 2002).

Based on these findings it is plausible that C1orf32 plays an important role in lipoproteic homeostasis in the retina and may be indirectly involved in AMD pathogenesis. Therefore a top priority would be to investigate whether LISCH7 is also expressed in retina or if its receptor role is taken over by C1orf32 in the retina. It would also be recommendable to perform functional studies such as the investigation of lipoproteic uptake and degradation in cells transfected with the C1orf32 transcript to test its role. A third investigative approach should be the identification of the cellular sources of C1orf32 production in retina to see if it co-localizes with ApoE immunoreactivity. Finally, a haplotype analysis of C1orf32 could be carried out in AMD patients and controls.

12. Characterization of DAPL1

Expression analysis by RT-PCR, virtual Northern blot, and qRT-PCR of a clone isolated from the retSSH library revealed the existence of a gene with very high expression in RPE and retina but almost inexistent expression in other neuronal tissues (Fig. 52). Based on this clone and other publicly available sequences, the 552 bp sequence of DAPL1 could be established. The virtual Northern blot analysis confirmed that the full-length sequence had been indeed attained as only one very strong signal of approximately 0.6 kb was seen in RPE and retina.

The 107 aa putative protein encoded by the gene does not contain any known motif, but has an endoplasmic reticulum (ER) retention signal (QPRK). Homology searches determined that the protein is highly similar to a number of death-associated proteins (DAP) from various species ranging from fruit fly to cow (Fig. 54). Whereas in humans until now only the sequence of one DAP protein has been reported, two DAP proteins are reported to exist in zebrafish (Inohara and Nunez 2000 – unpublished).

The nucleotide and protein sequence searches also led to the identification of 14 putative proteins similar to DAPL1. A multiple sequence alignment of these proteins revealed not only that the ER signal is present in all of them, but also the existence of a highly conserved 15 aa stretch in the N-terminal of the protein which is not reported in any of the motif or domain databases.

Given the high similarity and conservation of the protein sequences and the absence of clues about the physiological function of DAPL1 it is worthwhile to analyze what is known about DAP. DAP (also known as DAP1) was identified, together with the DAP-kinase (DAPK) gene in 1995 by Deiss et al., as mediator of interferon gamma (INF-?) induced cell death. Immunostaining and biochemical fractionation established that the protein is localized in the cytoplasm (Levy-Strumpf and Kimchi 1998) but in spite of the time since it has been cloned, the exact biochemical function of the protein remains unknown. Its role in apoptotic cell-death was independently confirmed in the Levy-Strumpf and Kimchi study (1998). They found that the death-promoting effects of the protein were prominent, so that over-expression of the full-length protein potentiated the killing effects of INF-?.

Apoptosis, or programmed cell death, plays a very important role in the homeostasis of a tissue. This mechanism is responsible for the elimination of damaged or malfunctioning cells. Although it has

widespread biological significance, being involved in embryogenesis, differentiation, proliferation, homoeostasis, removal of defect or harmful cells, and in regulation and function of the immune system, pathologic apoptosis can be associated with neurodegeneration, ischemia-reperfusion injury, bone marrow diseases, and cancer (Thompson 1995). Studies in the retina have shown that apoptosis is responsible for cell death in retinitis pigmentosa (Portera-Cailliau et al. 1994 and Reme et al. 1998) and it could be involved in the pathology of AMD (Xu et al. 1996 and Dunaief et al. 2002). It has been postulated that cell loss in AMD occurs by apoptosis of RPE cells, which is followed by death of the overlying photoreceptors, with rod cell loss preceding that of cones (Curcio 2001 and Dunaief et al. 2002).

Therefore, the finding of a gene that is expressed at very high levels exclusively in the RPE and retina, is highly conserved in several species, and has similarity to a gene known to be involved in apoptosis constitutes an interesting finding. Of course, there is no guarantee that sequence similarity between two proteins will result in similar biochemical function and in-depth studies of DAPL1 function should be performed to determine if DAPL1 is part of an apoptotic pathway.

13. Future goals

Although a number of the retina and neuronal genes identified in the course of this project have already been cloned, numerous other candidates still await to be characterized and this effort should be continued in the future. Certainly the greatest challenge and potential lies in the functional characterization of the cloned genes and the establishment of their role in the healthy retina as well as in retinal degenerations. Towards this aim, efforts are already underway to establish SNP maps of the genes in order to facilitate rapid investigation of their involvement in retinal disease.

VII References

Abalain JH, Carre JL, Leglise D, Robinet A, Legall F, Meskar A, Floch HH, Colin J. 2002. Is age-related macular degeneration associated with serum lipoprotein and lipoparticle levels? Clin Chim Acta 326:97-104

Adams MD. Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, et al. 1991. Complementary DNA sequencing: "expressed sequence tags" and the human genome project. Science 252:1651-1656

Ahmad I, Redmond LJ, Barnstable CJ, Adams MD, Kelley JM, Gocayne JD, Dubnick M. 1990. Developmental and tissue-specific expression of the rod photoreceptor cGMP-gated ion channel gene. Biochemical and Biophysical Research Communications 173:463-470

Allikmets R, Shroyer NF, Singh N, Seddon JM, Lewis RA, Bernstein PS, Peiffer A, Zabriskie NA, Li Y, Hutchinson A, Dean M, Lupski JR, Leppert M. 1997. Mutation of the Stargardt disease gene (ABCR) in age-related macular degeneration. Science 277:1805-1807

Allikmets R, Singh N, Sun H, Shroyer NF, Hutchinson A, Chidambaram A, Gerrard B, Baird L, Stauffer D, Peiffer A, Rattner A, Smallwood P, Li Y, Anderson KL, Lewis RA, Nathans J, Leppert M, Dean M, Lupski JR. 1997. A photoreceptor cell-specific ATP-binding transporter gene (ABCR) is mutated in recessive Stargardt macular dystrophy. Nature Genet 15:236-246

Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M. 2001. Genomewide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet 69:936-950

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403-410

Alward WL. 2003. Biomedicine. A new angle on ocular development. Science 299:1527-8

Amaratunga A, Abraham CR, Edwards RB, Sandell JH, Schreiber BM, Fine RE. 1996. Apolipoprotein E is synthesized in the retina by Muller glial cells, secreted into the vitreous, and rapidly transported into the optic nerve by retinal ganglion cells. J Biol Chem 271:5628-5632

Anderson DH, Ozaki S, Nealon M, Neitz J, Mullins RF, Hageman GS, Johnson LV. 2001. Cellular sources of apolipoprotein E in the human retina and retinal pigmented epithelium: implications for the process of drusen formation. Am J Ophthalmol 131:767-781

Arita M, Sato Y, Miyata A, Tanabe T, Takahashi E, Kayden HJ, Arai H, Inoue K. 1995. Human alpha-tocopherol transfer protein: cDNA cloning, expression and chromosomal localization. Biochemistry J 306:437-443

Armstrong B, Doll R. 1975. Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. Int J Cancer 15:617-631

Barnes WM. 1994. PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. Proc Natl Acad Sci USA 91:2216-2220

Bascom RA, Garcia-Heras J, Hsieh CL, Gerhard DS, Jones C, Francke U, Willard HF, Ledbetter DH, McInnes RR. 1992. Localization of the photoreceptor gene ROM1 to human chromosome 11 and mouse chromosome 19: sublocalization to human 11q13 between PGA and PYGM. American Journal Human Genetics 51:1028-1035

Bassler HA, Flood SJ, Livak KJ, Marmaro J, Knorr R, Batt CA. 1995. Use of a fluorogenic probe in a PCR-based assay for the detection of Listeria monocytogenes. Appl Environ Microbiol 61:3724-3728

Bavik CO, Levy F, Hellman U, Wernstedt C, Eriksson U. 1993. The retinal pigment epithelial membrane receptor for plasma retinol-binding protein. Journal Biological Chemistry 268:20540-20546

Bech-Hansen NT, Naylor MJ, Maybaum TA, Pearce WG, Koop B, Fishman GA, Mets M, Musarella MA, Boycott KM. 1998. Loss-of-function mutations in a calcium-channel alpha1-subunit gene in Xp11.23 cause incomplete X-linked congenital stationary night blindness. Nat Genet 19:264-267

Berger W, Meindl A, van de Pol TJR, Cremers FPM, Ropers HH, Doerner C, Monaco A, Bergen AAB, Lebo R, Warburg M, Zergollern L, Lorenz B, Gal A, Bleeker-Wagemakers EM, Meitinger T. 1992. Isolation of a candidate gene for Norrie disease by positional cloning. Nature Genet 1:199-203

Berget SM, Moore C, Sharp PA. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci USA 74:3171-3175

Berget SM. 1995. Exon recognition in vertebrate splicing. J Biol Chem 270:2411-2414

Bernstein SL, Wong P. 1998. Regional expression of disease-related genes in human and monkey retina. Molecular Vision 4:24

Bespalova IN, Buxbaum JD. 2003. Disease susceptibility genes for autism. Ann Med 35:274-281

Bihain BE, Yen FT.1998. The lipolysis stimulated receptor: a gene at last. Curr Opin Lipidol 9:221-224

Bishop JO, Morton JG, Rosbash M, Richardson. M. 1974. Three classes in Hela cell messenger RNA. Nature 250:199-204

Blackshaw S, Fraioli RE, Furukawa T, Cepko CL. 2001. Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes. Cell 107:579-589

Blackshaw S, Kuo WP, Park PJ, Tsujikawa M, Gunnersen JM, Scott HS, Boon WM, Tan SS, Cepko CL. 2003. MicroSAGE is highly representative and reproducible but reveals major differences in gene expression among samples obtained from similar tissues. Genome Biol 4:R17

Boguski MS, Lowe TM, Tolstoshev CM. 1993. dbEST--database for "expressed sequence tags". Nat Genet 4:332-333

Bonaldo MF, Lennon G, Soares MB. 1996. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res 6:791-806.

Bortoluzzi S, d'Alessi F, Danieli GA. 2000. A novel resource for the study of genes expressed in the adult human retina. Invest Ophthalmol Vis Sci 41:3305-3308

Bortoluzzi S, d'Alessi F, Romualdi C, Danieli GA. 2001. Differential expression of genes coding for ribosomal proteins in different human tissues. Bioinformatics 17:1152-1157

Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314-331

Bowmaker JK, Dartnall HJ. 1980. Visual pigments of rods and cones in a human retina. J Physiol 298:501-511

Bowne SJ, Daiger SP, Malone KA, Heckenlively JR, Kennan A, Humphries P, Hughbanks-Wheaton D, Birch DG, Liu Q, Pierce EA, Zuo J, Huang Q, Donovan DD, Sullivan LS. 2003. Characterization of RP1L1, a highly polymorphic paralog of the retinitis pigmentosa 1 (RP1) gene. Mol Vis 9:129-137

Brandstätter JH, Koulen P, Kuhn R, van der Putten H, Wassle H. 1996. Compartmental localization of a metabotropic glutamate receptor (mGluR7): two different active sites at a retinal synapse. J Neurosci 16:4749-4756

Burge CB and Karlin S. 1998. Finding the genes in genomic DNA. Curr Opin Struct Biol 8:346-354

Burge CB. 2001. Chipping away at the transcriptome. Nat Genet 27:232-234

Bustin SA. 2000. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. J Mol Endocrinol. 25:169-193

Cabot EL, Beckenbach AT. 1989. Simultanous editing of multiple nucleic acid and protein sequences with ESEE. Computer Applied Bioscience 5:233-234

Carninci P, Shiraki T, Mizuno Y, Muramatsu M, Hayashizaki Y. 2002. Extra-long first-strand cDNA synthesis. Biotechniques 32:984-985

Carninci P, Waki K, Shiraki T, Konno H, Shibata K, Itoh M, Aizawa K, Arakawa T, Ishii Y, Sasaki D, Bono H, et al. 2003. Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. Genome Res 13:1273-1289

Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285-298

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. Nat Genet 31:415-418

Chen ZY, Battinelli EM, Hendriks RW, Powell JF, Middleton-Price H, Sims KB, Breakefield XO, Craig IW. 1993. Norrie disease gene: characterization of deletions and possible function. Genomics 2:533-535

Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell 12:1-8

Christen WG, Glynn RJ, Manson JE, Ajani UA, Buring JE. 1996. A prospective study of cigarette smoking and age-related macular degeneration in men. JAMA 276:1147-1151

Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH, Johnston M. 2001. Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence. Genome Res 11:1175-1186

Collins A, Morton NE. 1998. Mapping a disease locus by allelic association. Proc Natl Acad Sci USA 95:1741-1745

Collins C, Hutchinson G, Kowbel D, Riess O, Weber B, Hayden MR. 1992. The human beta-subunit of rod photoreceptor cGMP phosphodiesterase: complete retinal cDNA sequence and evidence for expression in brain. Genomics 3:698-704

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA. 1993. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261:921-923

Couzin J. 2002. Breakthrough of the year. Small RNAs make big splash. Science 298:2296-2297

Crabb JW, Goldflam S, Harris SE, Saari JC. 1988. Cloning of the cDNAs encoding the cellular retinaldehyde-binding protein from bovine and human retina and comparison of the protein structuRes Journal Biological Chemistry 263:18688-186892

Craft CM, Whitmore DH, Donoso LA. 1990. Differential expression of mRNA and protein encoding retinal and pineal S-antigen during the light/dark cycle. Journal of Neurochemistry 55:1461-1473

Croft L, Schandorff S, Clark F, Burrage K, Arctander P, Mattick JS. 2000. ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. Nat Genet 24:340-341

Cruickshanks KJ, Klein R, Klein BEK. 1993. Sunlight and age related macular degeneration: the Beaver Dam Eye Study. Archives Ophthalmology 111:514-518

Culbertson MR.1999. RNA surveillance. Unforeseen consequences for gene expression, inherited genetic disorders and cancer. Trends Genet , 15:74-80

Curcio CA, Sloan KR, Kalina RE, Hendrickson AE. 1990. Human photoreceptor topography. J Comp Neurol 292:497-523

Curcio CA.. 2001. Photoreceptor topography in ageing and age-related maculopathy. Eye. 15:376-383

Dahlseid JN, Puziss J, Shirley RL, Atkin AL, Hieter P, Culbertson MR. 1998. Accumulation of mRNA coding for the ctf13p kinetochore subunit of Saccharomyces cerevisiae depends on the same factors that promote rapid decay of nonsense mRNAs. Genetics 150:1019-1035

Deiss LP, Feinstein E, Berissi H, Cohen O, Kimchi A. 1995. Identification of a novel serine/threonine kinase and a novel 15-kD protein as potential mediators of the gamma interferon-induced cell death. Genes Dev. 9:15-30

den Hollander AI, van Driel MA, de Kok YJ, van de Pol DJ, Hoyng CB, Brunner HG, Deutman AF, Cremers FP. 1999 . Isolation and mapping of novel candidate genes for retinal disorders using suppression subtractive hybridization. Genomics 58:240-249

Deutsch M, Long M. 1999. Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res 27:3219-3228

Dev KK, Nakanishi S, Henley JM. 2001. Regulation of mglu(7) receptors by proteins that interact with the intracellular C-terminus. Trends Pharmacol Sci 22:355-361

Diatchenko L, Lau YF, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov S, Lukyanov K, Gurskaya N, Sverdlov ED, Siebert PD. 1996. Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc Natl Acad Sci USA 93:6025-6030

Dockray GJ, Sault C, Holmes S. 1986. Antibodies to FMRF amide, and the related pentapeptide LPLRF amide, reveal two groups of immunoreactive peptides in chicken brain. Regul Pept. 16:27-37

Donders FC. 1855. Beiträge zur pathologischen Anatomie des Auges. Archives für Ophthalmologie 1:106-118

D'Onofrio C, Colantuoni V, Cortese R. 1985. Structure and cell-specific expression of a cloned human retinol binding protein gene: the 5'-flanking region contains hepatoma specific transcriptional signals. EMBO J 8:1981-1989

Dowling J. 1987. The Retina: An approachable part of the brain. Cambridge, MA: Belknap Press,

Dryja TP, Hahn LB, Reboul T, Arnaud B. 1996. Missense mutation in the gene encoding the alpha subunit of rod transducin in the Nougaret form of congenital stationary night blindness. Nature Genetics 13:358-360

Dryja TP, McGee TL, Hahn LB, Cowley GS, Olsson JE, Reichel E, Sandberg MA, Berson EL. 1990. Mutations within the rhodopsin gene in patients with autosomal dominant retinitis pigmentosa. New Eng. J. Med. 323:1302-1307

Dryja TP, McGee TL, Reichel E, Hahn LB, Cowley GS, Yandell DW, Sandberg MA, Berson EL. 1990. A point mutation of the rhodopsin gene in one form of retinitis pigmentosa. Nature 343:364-366

Duguid JR, Dinauer MC. 1990. Library subtraction of in vitro cDNA libraries to identify differentially expressed genes in scrapie infection. Nucleic Acids Res 18:2789-2792

Dunaief JL, Dentchev T, Ying GS, Milam AH. 2002. The role of apoptosis in age-related macular degeneration. Arch Ophthalmol 120:1435-1442

Eddy SR. 2001. Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2:919-929

Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. Trends Genet 19:362-365

Endege WO, Steinmann KE, Boardman LA, Thibodeau SN, Schlegel R. 1999. Representative cDNA libraries and their utility in gene expression profiling. Biotechniques. 26:542-548

Enz R, Croci C. 2003. Different binding motifs in metabotropic glutamate receptor type 7b for filamin A, protein phosphatase 1C, protein interacting with protein kinase C (PICK) 1 and syntenin allow the formation of multimeric protein complexes. Biochem J. 372:183-189

Enz R. 2002. The actin-binding protein Filamin-A interacts with the metabotropic glutamate receptor type 7. FEBS Lett. 514:184-188

Enz R. 2002. The metabotropic glutamate receptor mGluR7b binds to the catalytic gamma-subunit of protein phosphatase 1. J Neurochem 81:1130-1140

Ercolani L, Florence B, Denaro M, Alexander M. 1988. Isolation and complete sequence of a functional human glyceraldehyde-3-phosphate dehydrogenase gene. J Biol Chem 263:15335-15341

Ercolani L., Florence B., Denaro M., and Alexander M. 1988. Isolation and complete sequence of a functional human glyceraldehyde-3-phosphate dehydrogenase gene. J Biol Chem 263:15335-15341

Evans J, Wormald R. 1996. Is the incidence of registrable age-related macular degeneration increasing? Br J Ophthalmol 80:9-14

Evans JR. 2001. Risk factors for age-related macular degeneration. Prog Retin Eye Res 20:227-253

Feinstein E, Druck T, Kastury K, Berissi H, Goodart SA, Overhauser J, Kimchi A, Huebner K. 1995. Assignment of DAP1 and DAPK--genes that positively mediate programmed cell death triggered by IFN-gamma--to chromosome regions 5p12.2 and 9q34.1, respectively. Genomics. 29:305-307

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496-512

Flor PJ, Van Der Putten H, Ruegg D, Lukic S, Leonhardt T, Bence M, Sansig G, Knopfel T, Kuhn R. 1997. A novel splice variant of a metabotropic glutamate receptor, human mGluR7b. Neuropharmacology. 36:153-159

Fong SL 1992. Characterization of the human rod transducin alpha-subunit gene. Nucleic Acids Res 11:2865-2870

Franz O, Bruchhaus I, Roeder T. 1999. Verification of differential gene transcription using virtual northern blotting Nucleic Acids Res 27:e3

Frawley W, Piatetsky-Shapiro G and Matheus C. 1992. Knowledge Discovery in Databases: An Overview. AI Magazine Fall:213-228

Friedman DS, Katz J, Bressler NM, Rahmani B, Tielsch JM. 1999. Racial differences in the prevalence of age-related macular degeneration: the Blatimore Eye Survey. Ophthalmology 106:1049-1055

Frischmeyer PA, Dietz HC. 1999. Nonsense-mediated mRNA decay in health and disease. Hum Mol Genet 8:1893-1900

Fukusumi S, Habata Y, Yoshida H, Iijima N, Kawamata Y, Hosoya M, Fujii R, Hinuma S, Kitada C, Shintani Y, Suenaga M, Onda H, Nishimura O, Tanaka M, Ibata Y, Fujino M. 2001. Characteristics and distribution of endogenous RFamide-related peptide-1. Biochim Biophys Acta 1540:221-232

Furukawa T, Morrow EM, Cepko CL. 1997 Crx, a novel otx-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation. Cell 91:531-541

Furukawa T, Morrow EM, Cepko CL. 1997. Crx, a novel otx-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation. Cell 91:531-541

Gass JDM. 1997. *Stereoscopic atlas of macular diseases: diagnosis and treatment*. 4th edition. Mosby, London

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31:3784-3788

Goffeau A. 1997. DNA technology: Molecular fish on chips. Nature 385:202-203

Gordon W., and Bazan N. 1997. In *Biochemistry of the Eye*. Harding, J. (ed.), Chapman & Hall, London. 144-275

Granadino B, Gallardo ME, Lopez-Rios J, Sanz R, Ramos C, Ayuso C, Bovolenta P, Rodriguez de Cordoba S. 1999. Genomic cloning, structure, expression pattern, and chromosomal location of the human SIX3 gene. Genomics 55:100-105

Greenberg MJ, Price DA. 1992. Relationships among the FMRFamidelike peptides. Prog Brain Res 92:25-37

Gregory CY, Converse CA, Foulds WS. 1991. Specific radioimmunoassay to investigate rod outer segment phagocytosis by retinal pigment epithelium in vitro. Ophthalmic Research 23:171-176

Hackam AS, Bradford RL, Bakhru RN, Shah RM, Farkas R, Zack DJ, Adler R. 2003. Gene discovery in the embryonic chick retina. Mol Vision 9:262-276

Hammerle K, Shayan P, Niemeyer CM, Flotho C. 2003. Expression analysis of alpha-NAC and ANX2 in juvenile myelomonocytic leukemia using SMART polymerase chain reaction and "virtual Northern" hybridization. Cancer Genet CytoGenet 142:149-152

Heid CA, Stevens J, Livak KJ, Williams PM. 1996. Real time quantitative PCR. Genome Res 6:986-994

Higuchi R, Fockler C, Dollinger G, Watson R. 1993. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. Biotechnology (NY) 11:1026-1030

Hinuma S, Shintani Y, Fukusumi S, Iijima N, Matsumoto Y, Hosoya M, Fujii R, Watanabe T, Kikuchi K, Terao Y, Yano T, Yamamoto T, Kawamata Y, Habata Y, Asada M, Kitada C, Kurokawa T, Onda H, Nishimura O, Tanaka M, Ibata Y, Fujino M. 2000. New neuropeptides containing carboxy-terminal RFamide and their receptor in mammals. Nat Cell Biol 2:703-708

Hirbec H, Perestenko O, Nishimune A, Meyer G, Nakanishi S, Henley JM, Dev KK. 2002. PDZ proteins PICK1, GRIP, and syntenin bind multiple glutamate receptor subtypes. Analysis of PDZ binding motifs. J Biol Chem 277:15221-15224

Hosomi A, Goto K, Kondo H, Iwatsubo T, Yokota T, Ogawa M, Arita M, Aoki J, Arai H, Inoue K. 1998. Localization of alpha-tocopherol transfer protein in rat brain. Neuroscience Letters 256:159-162

Huang, X. 1994. On Global Sequence Alignment. Computer Applications in the Biosciences 10:227-235

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G. 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411:599-603

Inglehearn CF. 1998. Molecular genetics of human retinal dystrophies. Eye 12:571-579

Joensuu T, Hamalainen R, Yuan B, Johnson C, Tegelberg S, Gasparini P, Zelante L, Pirvola U, Pakarinen L, Lehesjoki AE, de la Chapelle A, Sankila EM. 2001. Mutations in a novel gene with transmembrane domains underlie Usher syndrome type 3. Am J Hum Genet 69:673-84

Jones RB, Wang F, Luo Y, Yu C, Jin C, Suzuki T, Kan M, McKeehan WL. 2001. The nonsense-mediated decay pathway and mutually exclusive expression of alternatively spliced FGFR2IIIb and -IIIc mRNAs. J Biol Chem 276:4158-4167

Kannan K, Amariglio N, Rechavi G, Jakob-Hirsch J, Kela I, Kaminski N, Getz G, Domany E, Givol D. 2001. DNA microarrays identification of primary and secondary target genes regulated by p53. Oncogene 20:2225-2234

Kanoh H, Kondoh H, Ono T. 1983. Diacylglycerol kinase from pig brain. Purification and phospholipid dependencies. J Biol Chem 258:1767-1774

Kaplan J. 2002. Genomics and medicine: hopes and challenges. Gene Ther 9:658-661

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. Science 296:916-919

Kato H, Tillotson J, Nichaman MZ, Rhoads GG, Hamilton HB. 1973. Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California. Am J Epidemiol 97:372-385

Katsanis N, Worley KC, Gonzalez G, Ansley SJ, Lupski JR. 2002. A computational/functional genomics approach for the enrichment of the retinal transcriptome and the identification of positional candidate retinopathy genes. Proc Natl Acad Sci USA 99:14326-14331

Kellner U. 1997. Hereditäre Netzhautdystrophien. Ophthalmologe 94:450-465

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D. 2002. The Human Genome Browser at UCSC. Genome Res 12:996-1006

Kikkawa Y, Shitara H, Wakana S, Kohara Y, Takada T, Okamoto M, Taya C, Kamiya K, Yoshikawa Y, Tokano H, Kitamura K, Shimizu K, Wakabayashi Y, Shiroishi T, Kominami R, Yonekawa H. 2003. Mutations in a new scaffold protein Sans cause deafness in Jackson shaker mice. Hum Mol Genet 12:453-456

Kinoshita A, Shigemoto R, Ohishi H, van der Putten H, Mizuno N. 1998. Immunohistochemical localization of metabotropic glutamate receptors, mGluR7a and mGluR7b, in the central nervous system of the adult rat and mouse: a light and electron microscopic study. J Comp Neurol. 393:332-352

Kitano H. 2002. Systems biology: a brief overview. Science 295:1662-1664

Klaver CC, Kliffen M, van Duijn CM, Hofman A, Cruts M, Grobbee DE, van Broeckhoven C, de Jong PT. 1998. Genetic association of apolipoprotein E with age-related macular degeneration. Am J Hum Genet 63:200-206

Klein ML, Schultz DW, Edwards A, Matise TC, Rust K, Berselli CB, Trzupek K, Weleber RG, Ott J, Wirtz MK, Acott TS. 1998. Age-related macular degeneration. Clinical features in a large family and linkage to chromosome 1q. Arch Ophthalmol 116:1082-1088

Klein R, Klein BE, Jensen SC, Meuer SM. 1997. The five-year incidence and progression of age-related maculopathy: the Beaver Dam Eye Study. Ophthalmology 104:7-21

Klein R, Rowland ML, Harris MI. 1995. Racial / ethnic differences in age-related maculopathy: Third National Health and Nutrition Survey. Ophthalmology 102:371-381

Kohl S, Marx T, Giddings I, Jägle H, Jacobson SG, Apfelstedt-Sylla E, Zrenner E, Sharpe LT, Wissinger B. 1998. Total colourblindness is caused by mutations in the gene encoding the alpha-subunit of the cone photoreceptor cGMP-gated cation channel. Nature Genetics 19:257-259

Krawczak M, Ball EV, Fenton I, Stenson PD, Abeysinghe S, Thomas N, Cooper DN. 2000. Human gene mutation database-a biomedical information and research resource. Hum Mutat. 15:45-51

Krawczak M, Reiss J, Cooper DN. 1992. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Hum Genet 90:41-54

Kremer H, Pinckers A, van den Helm B, Deutman AF, Ropers H-H, Mariman ECM. 1994. Localization of the gene for dominant cystoid macular dystrophy on chromosome 7p. Hum Mol Genet 3:299-302

Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, Sunyaev S. 2003. Increase of functional diversity by alternative splicing. Trends Genet 19:124-128

Krott R and Heimann K. 1996. Altersabhängige makuladegeneration. Deutsches Ärzteblatt 93:A-1039-A1042

Kuklin A, Munson K, Gjerde D, Haefele R, Taylor P. 1997-1998. Detection of single-nucleotide polymorphisms with the WAVE DNA fragment analysis system. Genet Test 1:201-206

Kumar, A. et al. 2002. An integrated approach for finding overlooked genes in yeast. Nature Biotechnol 20:58-63

Kuznetsov VA, Knott GD, Bonner RF. 2002. General statistics of stochastic process of gene expression in eukaryotic cells. Genetics 161:1321-1332

Labow BI, Souba WW, Abcouwer SF. 2001. Mechanisms governing the expression of the enzymes of glutamine metabolism--glutaminase and glutamine synthetase. J Nutr 131:2467S-2474S

Ladd AN, Cooper TA. 2002. Finding signals that regulate alternative splicing in the post-genomic era. Genome Biol 3:reviews0008

Lamba JK, Adachi M, Sun D, Tammur J, Schuetz EG, Allikmets R, Schuetz JD. 2003. Nonsense mediated decay downregulates conserved alternatively spliced ABCC4 transcripts bearing nonsense codons. Hum Mol Genet 12:99-109

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921

Lander ES. 1996. The new genomics: global views of biology. Science 274:536-539

Lee WH, Bookstein R, Hong F, Young LJ, Shew JY, Lee EY. 1987. Human retinoblastoma susceptibility gene: cloning, identification, and sequence. Science 235:1394-1399

Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, Parvizi B, Cheung F, Anotnescu V, White J, Holt I, Liang F, Quackenbush J. 2002. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). Genome Res 12:493-502

Lehninger AL, Nelson DL, Cox MM. 1982. In *Principles of Biochemistry*, 2nd edition, Worth Publishers, Worth, New York. 615-644

Levy-Strumpf N, Kimchi A. 1998. Death associated proteins (DAPs): from gene identification to the analysis of their apoptotic and tumor suppressive functions. Oncogene 17:3331-3340

Lewis BP, Green RE, Brenner SE. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci USA 100:189-192

Lin CT, Sargan DR. 2001. Generation and analysis of canine retinal ESTs: isolation and expression of retina-specific gene transcripts. Biochem Biophys Res Commun 282:394-403

Litt M, Luty JA. 1989. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44:397-401

Liu Q, Guan XM, Martin WJ, McDonald TP, Clements MK, Jiang Q, Zeng Z, Jacobson M, Williams DL Jr, Yu H, Bomford D, Figueroa D, Mallee J, Wang R, Evans J, Gould R, Austin CP. 2001. Identification and characterization of novel mammalian neuropeptide FF-like peptides that attenuate morphine-induced antinociception. J Biol Chem, 276:36961-36969

Lorenz W, Inglese J, Palczewski K, Onorato JJ, Caron MG, Lefkowitz RJ. 1991. The receptor kinase family: primary structure of rhodopsin kinase reveals similarities to the beta-adrenergic receptor kinase. Proceedings National Academy Science USA 88:8715-8719

Mackay TF. 2001. The genetic architecture of quantitative traits. Annu Rev Genet 35:303-339

Majewski J, Schultz DW, Weleber RG, Schain MB, Edwards AO, Matise TC, Acott TS, Ott J, Klein ML. 2003. Age-Related Macular Degeneration-a Genome Scan in Extended Families. Am J Hum Genet 73:540-550

Makoff A, Pilling C, Harrington K, Emson P. 1996. Human metabotropic glutamate receptor type 7: molecular cloning and mRNA distribution in the CNS. Brain Res Mol Brain Res 40:165-170

Malone K, Sohocki MM, Sullivan LS, Daiger SP. 1999. Identifying and mapping novel retinal-expressed ESTs from humans. Mol Vis 5:5

Maroni G. 1996. The organization of eukaryotic genes. Evol Biol 29:1-19

Marquardt A, Stohr H, Passmore LA, Kramer F, Rivera A, Weber BH. 1998. Mutations in a novel gene, VMD2, encoding a protein of unknown properties cause juvenile-onset vitelliform macular dystrophy (Best's disease). Hum Mol Genet 7:1517-525

Martin KJ, Pardee AB. 2000. Identifying expressed genes. Proc Natl Acad Sci USA 97:3789-3791

Masai I, Okazaki A, Hosoya T, Hotta Y. 1993. Drosophila retinal degeneration A gene encodes an eye-specific diacylglycerol kinase with cysteine-rich zinc-finger motifs and ankyrin repeats. Proc Natl Acad Sci USA 90:11157-11161

Masland RH. 2001. The fundamental plan of the retina. Nat Neurosci 4:877-886

Masugi M, Yokoi M, Shigemoto R, Muguruma K, Watanabe Y, Sansig G, van der Putten H, Nakanishi S. 1999. Metabotropic glutamate receptor subtype 7 ablation causes deficit in fear response and conditioned taste aversion. J NeuroSci 19:955-963

Mattick JS, Gagen MJ. 2001. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. Mol Biol Evol 18:1611-1630

Mattick JS. 2001. Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep. 2:986-989

Megy K, Audic S, Claverie JM. 2002. Heart-specific genes revealed by expressed sequence tag (EST) sampling. Genome Biol 3:RESEARCH0074

Meindl A, Dry K, Herrmann K, Manson F, Ciccodicola A, Edgar A, Carvalho MR, Achatz H, Hellebrand H, Lennon A, Migliaccio C, Porter K, Zrenner E, Bird A, Jay M, Lorenz B, Wittwer B, D'Urso M, Meitinger T, Wright A. 1996. A gene (RPGR) with homology to the RCC1 guanine nucleotide exchange factor is mutated in X-linked retinitis pigmentosa (RP3). Nature Genetics 13:35-42

Mendel G. 1866. Versuche über Pflanzenybridien. Verhandlungen des naturforschenden Vereins in Brünn. IV. Band. Abhandlungen 1865, Brünn. Im Verlage des Verein 3-47

Michelson AM, Marham AF; Orkin SH. 1983. Isolation and DNA sequence of a full-length cDNA clone for human X chromosome-encoded phosphoglycerate kinase. Proceedings National Academy of Science USA 80:472-476

Mitchell GA, Looney JE, Brody LC, Steel G, Suchanek M, Engelhardt JF, Willard HF, Valle D. 1988. Human ornithine-delta-aminotransferase. cDNA cloning and analysis of the structural gene. Journal Biological Chemistry 263:14288-14295

Mitchell P, Smith W, Wang JJ. 1998. Iris color, skin sun sensitivity, and age-related maculopathy: the Beaver Dam Eye Study. Ophthalmology 105:1359-1363

Modrek B, Lee CJ. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet 34:177-180

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

Modrek B. and Lee C. 2002. A genomic view of alternative splicing. Nat Genet 30:13-19

Mohler PJ, Gramolini AO, Bennett V. 2002. Ankyrins. J Cell Sci 115:1565-1566

Mosavi LK, Minor DL Jr, Peng ZY. 2002. Consensus-derived structural determinants of the ankyrin repeat motif. Proc Natl Acad Sci USA 99:16029-16034

Mu X, Zhao S, Pershad R, Hsieh TF, Scarpa A, Wang SW, White RA, Beremand PD, Thomas TL, Gan L, Klein WH. 2001. Gene expression in the developing mouse retina by EST sequencing and microarray analysis. Nucleic Acids Res 29:4983-4993

Nakajima Y, Yamamoto T, Nakayama T, Nakanishi S. 1999. A relationship between protein kinase C phosphorylation and calmodulin binding to the metabotropic glutamate receptor subtype 7. J Biol Chem 274:27573-27577

Nathans J, Hogness DS. 1984. Isolation and nucleotide sequence of the gene encoding human rhodopsin. Proc. Nat. Acad. Sci 81:4851-4855

Nathans J, Piantanida TP, Eddy RL, Shows TB, Hogness DS. 1986. Molecular genetics of inherited variation in human color vision. Science 232:203-210

North MA, Naggert JK, Yan Y, Noben-Trauth K, Nishina PM. 1997. Molecular characterization of TUB, TULP1, and TULP2, members of the novel tubby gene family and their possible relation to ocular diseases. Proceedings National Academy Science USA 94:3128-3133

Oduol F, Xu J, Niare O, Natarajan R, Vernick KD. 2000. Genes identified by an expression screen of the vector mosquito Anopheles gambiae display differential molecular immune response to malaria parasites and bacteria. Proc Natl Acad Sci USA 97:11397-11402

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, Achkar JP, Brant SR, Bayless TM, Kirschner BS, Hanauer SB, Nunez G, Cho JH. 2001. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 411:603-606

Okamoto N, Hori S, Akazawa C, Hayashi Y, Shigemoto R, Mizuno N, and Nakanishi S 1994. Molecular characterization of a new metabotorpic glutamate receptor mGluR7 coupled to inhibitory cyclic AMP signal transduction. J Biol Chem 269:1231-1236

Okubo, K, Hori, N, Matoba, R, Niiyama, T, Fukushima, A, Kojima, Y, & Matsubara, K: 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat Genet 2:173-179

Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J, Sussman JL, Verschueren KHG, Goldman A. 1992. The alpha/beta hydrolase fold. Protein Eng 5:197-211

Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T. 1989. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc Natl Acad Sci USA 86:2766-2770

Peltonen L, McKusick VA. 2001. Genomics and medicine. Dissecting human disease in the postgenomic era. Science 291:1224-1229

Penotti FE. 1991. Human pre-mRNA splicing signals. J Theor Biol 150:385-420

Perroy J, El Far O, Bertaso F, Pin JP, Betz H, Bockaert J, Fagni L. 2002. PICK1 is required for the control of synaptic transmission by the metabotropic glutamate receptor 7. EMBO J. 21:2990-2999

Petrukhin K, Koisti MJ, Bakall B, Li W, Xie G, Marknell T, Sandgren O, Forsman K, Holmgren G, Andreasson S, Vujic M, Bergen AAB, McGarty-Dugan V, Figueroa D, Austin CP, MetzkerL ML, Caskey CT, Wadelius C. 1998. Identification of the gene responsible for Best macular dystrophy. Nature Genetics 19:241-247

Pfaffl MW. 2001. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res 29:e45

Phelan JK, Bok D. 2002. A brief review of retinitis pigmentosa and the identified retinitis pigmentosa genes. Mol Vis 6:116-124

Pierce EA, Quinn T, Meehan T, McGee TL, Berson EL, Dryja TP. 1999. Mutations in a gene encoding a new oxygen-regulated photoreceptor protein cause dominant retinitis pigmentosa. Nature Genetics 22:248-254

Pinckers A, Deutman AF, Lion F, Aan de Kerk AL. 1983. Dominant cystoid macular dystrophy (DCMD). Ophthal Paediat Genet 3:157-167

Pittler SJ, Lee AK, Altherr MR, Howard TA, Seldin MF, Hurwitz RL, Wasmuth JJ, Baehr W. 1992. Primary structure and chromosomal localization of human and mouse rod photoreceptor cGMP-gated cation channel. Journal of biol chem 267:6257-6262

Portera-Cailliau C, Sung CH, Nathans J, Adler R. 1994. Apoptotic photoreceptor cell death in mouse models of retinitis pigmentosa. Proc Natl Acad Sci USA 91:974-978

Price DA, Greenberg MJ. 1977. Structure of a molluscan cardioexcitatory neuropeptide. Science. 197:670-671

Pritchard JK, Cox NJ. 2002. The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet 11:2417-2423

Qiu P, Benbow L, Liu S, Greene JR, Wang L. 2002. Analysis of a human brain transcriptome map. BMC Genomics 3:10

Raffa RB. 1988. The action of FMRFamide (Phe-Met-Arg-Phe-NH2) and related peptides on mammals. Peptides 9:915-22

Raffai RL, Dong LM, Farese RV Jr, Weisgraber KH. 2001. Introduction of human apolipoprotein E4 "domain interaction" into mouse apolipoprotein E. Proc Natl Acad Sci USA 98:11587-11591

Ramakers C, Ruijter JM, Deprez RH, Moorman AF. 2003. Assumption-free analysis of quantitative real-time PCR data. Neurosci Lett. 339:62-66

Rattner A, Sun H, Nathans J. 1999. Molecular genetics of human retinal disease. Annu Rev Genet 33:89-113

Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. 1997. GeneCards: integrating information about genes, proteins and diseases. Trends in Genetics 13:163

Reme CE, Grimm C, Hafezi F, Marti A, Wenzel A. 1998. Apoptotic cell death in retinal degenerations. Prog Retin Eye Res 17:443-464

Rezvani M, Barrans JD, Dai KS, Liew CC. 2000. Apoptosis-related genes expressed in cardiovascular development and disease: an EST approach. Cardiovasc Res 45:621-629

Risch NJ. 200. Searching for genetic determinants in the new millenium. Nature 405:847-856

Rogic S, Ouellette BF, Mackworth AK. 2002. Improving gene recognition accuracy by combining predictions from two gene-finding programs. Bioinformatics 18:1034-1045

Romualdi C, Bortoluzzi S, Danieli GA. 2001. Detecting differentially expressed genes in multiple tag sampling experiments: comparative evaluation of statistical tests. Hum Mol Genet 10:2133-2141

Rosenfeld PJ, Cowley GS, McGee TL, Sandberg MA, Berson EL, Dryja TP. 1992. A null mutation in the rhodopsin gene causes rod photoreceptor dysfunction and autosomal recessive retinitis pigmentosa. Nat Genet 1:209-213

Rychlik W, Rhoads RE. 1989. A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. Nucleic Acids Res 17:8543-8551

Sansig G, Bushell TJ, Clarke VR, Rozov A, Burnashev N, Portet C, Gasparini F, Schmutz M, Klebs K, Shigemoto R, Flor PJ, Kuhn R, Knoepfel T, Schroeder M, Hampson DR, Collett VJ, Zhang C, Duvoisin RM, Collingridge GL, van Der Putten H. 2001. Increased seizure susceptibility in mice lacking metabotropic glutamate receptor 7. J NeuroSci 21:8734-8745

Sauer C. 2001. *Untersuchungen zu den genetischen Ursachen hereditärer Netzhautdegenerationen des Menschen: Die Positionsklonierung als Strategie zur Isolierung und Charakterisierung retinaler Gene.* 79-80

Sauer CG, Gehrig A, Warneke-Wittstock R, Marquardt A, Ewing CC, Gibson A, Lorenz B, Jurklies B, Weber BHF. 1997. Positional cloning of the gene associated with X-linked juvenile retinoschisis. Nature Genetics 17:164-170

Saugstad JA, Kinzie JM, Mulvihill ER, Segerson TP, Westbrook GL. 1994. Cloning and expression of a new member of the L-2-amino-4-phosphonobutyric acid-sensitive class of metabotropic glutamate receptors. Mol Pharmacol 45:367-372

Schachat AP, Hyman L, Leske MC, Connell AM, Wu SY. 1995. Features of age-related macular degeneration in a black population. The Barbados Eye Study Group. Archives of Ophthalmology 113:728-735

Schick JH, Iyengar SK, Elston RC, Fijal BA, Klein BE, Klein R. 2001. The genetic epidemiology of age-related maculopathy IJHG 1:11-24

Schmidt S, Klaver C, Saunders A, Postel E, De La Paz M, Agarwal A, Small K, Udar N, Ong J, Chalukya M, Nesburn A, Kenney C, Domurath R, Hogan M, Mah T, Conley Y, Ferrell R, Weeks D, de Jong PT, van Duijn C, Haines J, Pericak-Vance M, Gorin M. 2002. A pooled case-control study of the apolipoprotein E (APOE) gene in age-related maculopathy. Ophthalmic Genet 23:209-223

Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL. 2000. Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity.Cell 101:671-684

Schoepp DD. 2001. Unveiling the functions of presynaptic metabotropic glutamate receptors in the central nervous system. J Pharmacol Exp Ther. 299:12-20

Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Hudson TJ, et al. 1996. A gene map of the human genome. Science 274:540-546

Schultz J, Doerks T, Ponting CP, Copley RR, Bork P. 2000. More than 1,000 putative new human signalling proteins revealed by EST data mining. Nat Genet 25:201-204

Schwahn U, Lenzner S, Dong J, Feil S, Hinzmann B, van Duijnhoven G, Kirschner R, Hemberger M, Bergen AAB, Rosenberg T, Pinckers AJLG, Fundele R, Rosenthal A, Cremers FPM, Ropers HH, Berger W. 1998. Positional cloning of the gene for X-linked retinitis pigmentosa 2. Nature Genetics 19:327-332

Seddon JM, Willet WC, Speizer FE, Hankinson SE. 1996. A prospective study of cigarette smoking and age-related macular degeneration in women. JAMA 276:1141-1146

Seddon JM. 2001. Epidemiology of age-related macular degeneration. In: Schachat AP, Ryan SJ, eds. *Retina*. 3rd ed. St Louis, Mo: Mosby 1039-1050

Sedgwick SG, Smerdon SJ. 1999. The ankyrin repeat: a diversity of interactions on a common structural framework.Trends Biochem Sci 24:311-316

Shabalina SA, Ogurtsov AY, Kondrashov VA, Kondrashov AS. 2001. Selective constraint in intergenic regions of human and mouse genomes. Trends Genet 17:373-376

Sharma S, Chang JT, Della NG, Campochiaro PA, Zack DJ. 2002. Identification of novel bovine RPE and retinal genes by subtractive hybridization. Mol Vis 8:251-258

Sharon D, Blackshaw S, Cepko CL, Dryja TP. 2002. Profile of the genes expressed in the human peripheral retina, macula, and retinal pigment epithelium determined through serial analysis of gene expression (SAGE). Proc Natl Acad Sci USA 99:315-320

Sharp PA. 1994. Split genes and RNA splicing. Cell 77:805-815

Shimizu-Matsumoto A, Adachi W, Mizuno K, Inazawa J, Nishida K, Kinoshita S, Matsubara K, Okubo K. 1997. An expression profile of genes in human retina and isolation of a complementary DNA for a novel rod photoreceptor protein. Invest Ophthalmol Vis Sci 38:2576-2585

Shyjan AW, de Sauvage FJ, Gillett NA, Goeddel DV, Lowe DG. 1992. Molecular cloning of a retina-specific membrane guanylyl cyclase. Neuron 4:727-737

Silbiger SM, Jacobsen VL, Cupples RL, Koski RA. 1994. Cloning of cDNAs encoding human TIMP-3, a novel member of the tissue inhibitor of metalloproteinase family. Gene 141:293-297

Simonelli F, Margaglione M, Testa F, Cappucci G, Manitto MP, Brancato R, Rinaldi E. 2001. E polymorphisms in age-related macular degeneration in an Italian population.Ophthalmic Res 33(6):325-328

Sinha S, Sharma A, Agarwal N, Swaroop A, Yang-Feng TL. 2000. Expression profile and chromosomal location of cDNA clones, identified from an enriched adult retina library. Invest Ophthalmol Vis Sci 41:24-28

Skerry TM, Genever PG. 2001. Glutamate signalling in non-neuronal tissues. Trends Pharmacol Sci 22:74-81

Sohocki MM, Malone KA, Sullivan LS, Daiger SP. 1999. Localization of retina/pineal-expressed sequences: identification of novel candidate genes for inherited retinal disorders. Genomics 58:29-33

Sorek R, Safer HM. 2003. A novel algorithm for computational identification of contaminated EST libraries. Nucleic Acids Res 31:1067-1074

Souied EH, Benlian P, Amouyel P, Feingold J, Lagarde JP, Munnich A, Kaplan J, Coscas G, Soubrane G. 1998. The epsilon4 allele of the apolipoprotein E gene as a potential protective factor for exudative age-related macular degeneration. Am J Ophthalmol 125:353-359

Souied EH, Ducroq D, Rozet JM, Gerber S, Perrault I, Munnich A, Coscas G, Soubrane G, Kaplan J. 2000. ABCR gene analysis in familial exudative age-related macular degeneration. Invest Ophthalmol Vis Sci 41:244-247

Spalton DJ, Hitchings RA, Holder GE. 1993. In *Atlas of clinical Ophthalmology*, 2nd edition, Mosby-Year Book Europe Limited, London

Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O, Zhang MQ. 2000. An alternative-exon database and its statistical analysis. DNA Cell Biol 19:739-756

Staudinger J, Lu J, Olson EN. 1997. Specific interaction of the PDZ domain protein PICK1 with the COOH terminus of protein kinase C-alpha. J Biol Chem 272:32019-32024

Stern MD, Anisimov SV, Boheler KR. 2003. Can transcriptome size be estimated from SAGE catalogs? Bioinformatics 19:443-448

Stöhr H, Mah N, Schulz H, Gehrig A, Frohlich S, Weber B. 2000. EST mining of the UniGene dataset to identify retina-specific genes. Cytogenet Cell Genet 91:267-277

Stöhr H. 2003. Die Genetik der komplexen AMD. Medgen 15:117-123

Stone EM, Lotery AJ, Munier FL, Héon E, Piguet B, Guymer RH, Vandenburgh K, Cousin P, Nishimura D, Swiderski RE, Silvestri G, Mackey DA, Hageman GS, Bird AC, Sheffield VC, Schorderet DF. 1999. A single EFEMP1 mutation associated with both Malattia Leventinese and Doyne honeycomb retinal dystrophy. Nature Genetics 22:199-202

Stone EM, LoteryAJ, MunierFL, Heon E, Piguet B, Guymer RH, Vandenburgh K, Cousin P, Nishimura D, Swiderski RE, Silvestri G, Mackey DA, Hagerman GS, Bird AC, Sheffield VC, Schorderet DF. 1999. A single EFEMP1 mutation associated with both malattia Leventinese and Doyne honeycomb retinal dystrophy. Nature Genet 22:199-202

Stormo GD. 2000. Gene-finding approaches for eukaryotes. Genome Res 10:394-397

Stowell JN and Craig AM. 1999. Axon/dendrite targeting of metabotropic glutamate receptors by their cytoplasmic carboxy-terminal domains. Neuron 22:525–536

Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB. 2002. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci USA 99:4465-4470

Suzuki T, Higgins PJ, Crawford DR: 2000. Control selection for RNA quantitation. Biotechniques 29:332-337

Swaroop A, Xu JZ, Pawar H, Jackson A, Skolnick C, Agarwal N. 1992. A conserved retina-specific gene encodes a basic motif/leucine zipper domain. Proceedings of the National Academy of Science USA 89:266-270

Swaroop A and zack DJ. 2002. Transcriptome analysis of the retina. Genome Biol 3:reviews1022.1-1022.4

Szymanski M, Barciszewski J. 2002. Beyond the proteome: non-coding regulatory RNAs. Genome Biol 3:reviews0005

Tabor HK, Risch NJ, Myers RM. 2002. Opinion: Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 3:391-397

Terzic J, Muller C, Gajovic S, Saraga Babic M. 1998. Expression of PAX2 gene during human development. International Journal of Developmental Biology 5:701-707

Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, Grisar T, Igout A, Heinen E: 1999. Housekeeping genes as internal standards: use and limits. J Biotechnol 75:291-295

Thompson CB.1995. Apoptosis in the pathogenesis and treatment of disease. Science 267:1456-1462

Tichopad A, Dzidic A, Pfaffl MW. 2002. Improving quantitative real-time RT-PCR reproducibility by boosting primer-linked amplification efficiency. Biotechnology Letters 24:2053-2056

Travis GH, Sutcliffe JG, Bok D. 1991. The retinal degeneration slow (rds) gene product is a photoreceptor disc membrane-associated glycoprotein. Neuron 1:61-70

Tricarico C, Pinzani P, Bianchi S, Paglierani M, Distante V, Pazzagli M, Bustin SA, Orlando C. 2002. Quantitative real-time reverse transcription polymerase chain reaction: normalization to rRNA or single housekeeping genes is inappropriate for human tissue biopsies. Anal BioChem 309:293-300

Tuteja N, Danciger M, Klisak I, Tuteja R, Inana G, Mohandas T, Sparkes RS, Farber DB. 1990. Isolation and characterization of cDNA enconding the gamma-subunit of cGMP phosphodiesterase in human retina. Gene 88:227-232

Vanden Langenberg GM, Mares-Perlman JA, Klein R, Klein BE, Brady WE, Palta M. 1998. Associations between antioxidant and zinc intake and the 5-year incidence of early age-related maculopathy in the Beaver Dam Eye Study. American Journal of Epidemiology 148:204-214

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 3:RESEARCH0034

Vasmatzis G, Essand M, Brinkmann U, Lee B, Pastan I. 1998. Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. Proc Natl Acad Sci USA 95:300-304

Velculescu VE, Madden SL, Zhang L, Lash AE, Yu J, Rago C, Lal A, Wang CJ, Beaudry GA, Ciriello KM, Cook BP, Dufault MR, Ferguson AT, Gao Y, He TC, Hermeking H, Hiraldo SK, Hwang PM, Lopez MA, Luderer HF, Mathews B, Petroziello JM, Polyak K, Zawel L, Kinzler KW, et al. Analysis of human transcriptomes. 1999. Nat Genet 23:387-388

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, et al. 2001. The sequence of the human genome. Science 291:1304-1351

Versteeg R, Van Schaik BD, Van Batenburg MF, Roos M, Monajemi R, Caron H, Bussemaker HJ, Van Kampen AH. 2003. The Human Transcriptome Map Reveals Extremes in Gene Density, Intron Length, GC Content, and Repeat Pattern for Domains of Highly and Weakly Expressed Genes. Genome Res Aug 12 [Epub ahead of print]

Vollrath L. 1985. Mammalian pinealocytes: Ultrastructural aspects and innervation. In “Photoperiodism, melatonin and the pineal”, Pitman. Avon, UK

Wang J, Chai X, Eriksson U, Napoli JL. 1999. Activity of human 11-cis-retinol dehydrogenase (Rdh5) with steroids and retinoids and expression of its mRNA in extra-ocular human tissue. Biochemistry Journal 338:23-27

Wang Q, Chen Q, Zhao K, Wang L, Wang L, Traboulsi EI. 2001. Update on the molecular genetics of retinitis pigmentosa. Ophthalmic Genet 22:133-154

Weber JL, May PE. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388-396

Weeks DE, Conley YP, Tsai HJ, Mah TS, Rosenfeld PJ, Paul TO, Eller AW, Morse LS, Dailey JP, Ferrell RE, Gorin MB. 2001. Age-related maculopathy: an expanded genome-wide scan with evidence of susceptibility loci within the 1q31 and 17q25 regions. Am J Ophthalmol 132:682-692

Weeks DE, Lathrop GM. 1995. Polygenic disease: methods for mapping complex disease traits. Trends Genet 11:513-519

Weil D, El-Amraoui A, Masmoudi S, Mustapha M, Kikkawa Y, Laine S, Delmaghani S, Adato A, Nadifi S, Zina ZB, Hamel C, Gal A, Ayadi H, Yonekawa H, Petit C. 2003. Usher syndrome type I G (USH1G) is caused by mutations in the gene encoding SANS, a protein that associates with the USH1C protein, harmonin. Hum Mol Genet 12:463-471

Weleber RG, Carr RE, Murphey WH, Sheffield VC, Stone EM. 1993. Phenotypic variation including retinitis pigmentosa, pattern dystrophy, and fundus flavimaculatus in a single family with a deletion of codon 153 or 154 of the peripherin/RDS gene. Arch Ophthalmol 111:1531-1542

Wilda M, Bachner D, Zechner U, Kehrer-Sawatzki H, Vogel W, Hameister H. 2000. Do the constraints of human speciation cause expression of the same set of genes in brain, testis, and placenta? Cytogenet Cell Genet 91:300-302

Wilson EB. 1911. The sex chromosomes. Arch Mikrosk Anat Entwiclungsmech 77:249-271

Winkelmann BR, Hager J, Kraus WE, Merlini P, Keavney B, Grant PJ, Muhlestein JB, Granger CB. 2000. Genetics of coronary heart disease: current knowledge and research principles. Am Heart J 140:11-26

Wistow G, Bernstein SL, Wyatt MK, Ray S, Behal A, Touchman JW, Bouffard G, Smith D, Peterson K. 2002. Expressed sequence tag analysis of human retina for the NEIBank Project: retbindin, an abundant, novel retinal cDNA and alternative splicing of other retina-preferred gene transcripts. Mol Vis 8:196-204

Wolfe KH, Li WH. 2003. Molecular evolution meets the genomics revolution. Nat Genet 33:255-265

Wolgemuth DJ, Watrin F. 1991. List of cloned mouse genes with unique expression patterns during spermatogenesis. Mamm Genome 1:283-288

Wood AJJ. 2000. Age-related macular degeneration. The New England Journal of Medicine 432:483-492

Xu GZ, Li WW, Tso MO. 1996. Apoptosis in human retinal degenerations. Trans Am Ophthalmol Soc. 94:411-430

Xu Q, Modrek B, Lee C. 2002. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res 30:3754-3766

Yates JRW, Moore AT. 2000. Genetic susceptibility to age related macular degeneration. Journal of Medical Genetics 37:83-87

Yen FT, Mann CJ, Guermani LM, Hannouche NF, Hubert N, Hornick CA, Bordeau VN, Agnani G, Bihain BE. 1994. Identification of a lipolysis-stimulated receptor that is distinct from the LDL receptor and the LDL receptor-related protein. Biochemistry 33:1172-1180

Yen FT, Masson M, Clossais-Besnard N, Andre P, Grosset JM, Bougueleret L, Dumas JB, Guerassimenko O, Bihain BE. 1999. Molecular cloning of a lipolysis-stimulated remnant receptor expressed in the liver. J Biol Chem 274:13390-13398

Yoshida H, Habata Y, Hosoya M, Kawamata Y, Kitada C, Hinuma S. 2003. Molecular properties of endogenous RFamide-related peptide-3 and its interaction with receptors. Biochim Biophys Acta 1593:151-157

Young Je, Vogt T, Gross KW, Khani SC. 2003. A short, highly active photoreceptor-specific enhancer/promoter region upstream of the human rhodopsin kinase gene. Invest Ophthalmol Vis Sci 44:4076-4085

Zarbin M. 1998. Age-related macular degeneration: review of pathogenesis. European Journal of Ophthalmology 8:199-206

Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, Hayashizaki Y, Gaasterland T; RIKEN GER Group; GSL Members. 2003. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Res 13:1290-1300

Zavolan M, van Nimwegen E, Gaasterland T. 2002. Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome. Genome Res 12:1377-1385

Zhang MQ. 2002. Computational prediction of eukaryotic protein-coding genes. Nat Rev Genet 9:698-709

Zhang Y, Mei P, Lou R, Zhang MQ, Wu G, Qiang B, Zhang Z, Shen Y. 2002. Gene expression profiling in developing human hippocampus. J Neurosci Res 70:200-208

Zhao X, Huang J, Khani SC, Palczewski K. 1998. Molecular forms of human rhodopsin kinase (GRK1). J Biol Chem 273:5124-5131

VIII Appendix

1. Primers

Table 35 General primers

Primer name	Sequence (5'? 3')
3'-RACE AP	GGCCACGCGTCGACTAGTACT(TTT)$_8$(AGC)
5' PCR primer	AAGCAGTGGTATCAACGCAGAGT
5' RACE inner primer	CGCGGATCCGAACACTGCGTTTGCTGGCTTTGATG
5' RACE outer primer	GCTGATGGCGATGAATGAACACTG
AAP	GGCCACGCGTCGACTAGTACGGGIIGGGIIGGGIIG
AP1	CCATCCTAATACGACTCACTATAGGGC
AP2	ACTCACTATAGGGCTCGAGCGGC
AUAP	GGCCACGCGTCGACTAGTAC
CDSIII/3' PCR primer	ATTCTAGAGGCCGAGGCGGCCGACATG-d(T)30N–1N
G3PDH_F	ATCGTGGAAGGACTCATGACC
G3PDH_F1	TGCACCACCAACTGCTTAGC
G3PDH_R	AGCGCCAGTAGAGGCAGGGAT
G3PDH_R1	GGCATGGACTGTGGTCATGAG
GUSB3	ACTATCGCCATCAACAACACACTGACC
GUSB5	GTGACGGTGATGTCATCGAT
GUSB6	GATCCACCTCTGATGTTCAC
GUSB7	CCTTTAGTGTTCCCTGCTAG
M13F	CGCCAGGGTTTTCCCAGTCACGAC
M13R	AGCGGATAACAATTTCACACAGGA
PCR-f	CTCGGATCCACTAGTAACGG
PCR-r	GCCGCCAGTGTGATGGATAT
?gt10F	GGTGGCGACTCCTGGAGCCCG
?gt10R	TTGACACCAGACCAACTGGTAATG
?gt11F (EKLC39)	GGTGGCGACGACTCCTGGAGCCCG
?gtR (EKLC40)	TTGACACCAGACCAACTGGTAATG
?TriplEx2 3'	CTCACTATAGGGCGAATTGGC
?TriplEx2 5'	CTCGGGAAGCGCGCCATTGT

Table 36 Primers used for RT-PCR

Primer name	Sequence (5' ? 3')	Primer name	Sequence (5' ? 3')	Primer name	Sequence (5' ? 3')
A001aR	CCACGAAGACTGGGATTT	**A020R**	TCCTAACCCCAGCTCCTC	**A38F**	TAGGCAGAGGTGGATGGG
A001F	TGTCTGTTAGTGCTCTCC	**A021F**	TCAACTCCTGCCTTCTTC	**A38F1**	TCAGCCCCGAGGTGTTTGC
A002F	TGTGTTGTGCATTCGTGC	**A021R**	GCCAGTGAGGACAGGAAC	**A38F2**	GACTGCCTCATGATACCC
A002R	AAGCACTGGTAACTGGCG	**A022aF**	AAGCTCTCCTCCCAGAAG	**A38F3**	CGGAACCGCTGTGAGTGC
A003aF	TGGCTGGCGTGACCTTGG	**A022aR**	CTTTTCTCAGCCCATCCG	**A38F4**	CATGCTACCACGGCTTCC
A003aR	GGCAGGGAAGGAGGCGGT	**A022F**	TCTCCAGCATTAAGCAGC	**A38F5**	GTGCTGACCGACAGCTTC
A004F	GCCACCTAAGCACAGTTCC	**A022R**	CGCTTTGTGCTTGGATTC	**A38F5**	GTGCTGACCGACAGCTTC
A004F	GCCACCTAAGCACAGTTCC	**A023F**	AGCGCGAGGAGAACTCAG	**A38F6**	CCAGGGATAGCACCAGCC
A004R	AAAATCCCTGTATCTGCC	**A023R**	CCGGCTTTTCTCCACAGG	**A38F8**	CCCTCTCTGGACCTAAGTG
A005F	TCCCTCTCTTGCTTTGGC	**A024F**	CTCTAACTCCGTGGCTGC	**A38R**	TGCCAAGCTGTTAGTGCC
A005R	AAGAATGCCCTCCTGCTC	**A024R**	CAGTTTCGGCTGCATTTC	**A38R2**	GTATGCTCTGTATGGATGG
A006F	TCATGGAACACCGCCCTC	**A025F**	GGGCCGACACAATTCACG	**A38R3**	GGCCACTCGGGCTTGTAG
A006R	CGCAAACGCCAAAATGTG	**A025R**	GCACGTCTGGGGCATTTC	**A38R4**	GTGCAATGCCAGCTCTTC
A007F	TGGTGCATACCTCAAACC	**A026F**	AAGGCGCAAAGTCACAGG	**A38R5**	CCCTCCTTTGTCCTGCTG
A007R	GGTCTTTGCTGCATTCTC	**A026R**	GCTCTCACCCACCTGCAC	**A38R6**	GAATATAAACAGCTTCCTTAGTC
A008F	TGCCCATCAGGACAAAAG	**A027F**	ATGCAAAATGATGGAAGG	**A38rR7**	TAGAGGCACGTGGCCTGCTT
A008R	TTCTGTTTTGTTGTGGGC	**A027R**	GGCCATAAAGCACAACTC	**A039F**	TGCATTGATTCTGCTTAGC
A009F	TCTGAAGTGGGAGTTGGG	**A028F**	CCTTGGCACCAACATGGA	**A039R**	TGCATTGATTCTGCTTAGC
A009R	AGCTTCTGGGTCAGATGG	**A028R**	GGAGCCGCCTACTTTTTG	**A040F**	CCAAGACCAGCTCACGAC
A010F	GGGAAGCTCTGAAGGCAG	**A029F**	ATCATACCACACTCGTCC	**A040R**	ATCCCACCCTCCTCAGTC
A010R	TTAGGCGGGGAAAAGGTC	**A029R**	GGGCAAAAAAAGGACTAC	**A042F2**	AAATAAAGCAGGGCACGC
A011F	AAATGGAATAGGAAAGGC	**A030F**	CCCATTCTCTCTTCGTGC	**A042R2**	CCTTTGGCCTTTCTTCAG
A011R	GCGGTTTTCATATTTTGC	**A030R**	CCTTCATTGGGCTTCATC	**A043R**	TACTTTCCCTTTGCTGCC
A012F	GTACCCAAGAGAAAGCCC	**A031F**	CTTGATACGCAGTTTAGTC	**A044F**	GCTGTCCATGCAGAACAC
A012R	GGGATGTAAAATTGGGATG	**A031R**	AAGCAAGTTTATCTATGGG	**A044R**	TTTGGAGCCCATCTAGTC
A013F	AAAGAGAACCTGCCCACA	**A032R**	GTCGTATTTGCTGTGCCC	**A045F**	TCAGGTAAACTGGGCCTC
A013R	AGAAGGCAGGGTATAGGG	**A033F**	TTGGGAGAAAGAGGCCAG	**A045R**	AGACACATGATGGCCCAG
A014F	TCAGTGAAAGTGGTGGCC	**A033R**	CTGTCCATGACTCAGGGC	**A046F**	CTAGGGCTGTGACAATGG
A014R	TTGTGTTCTCCCGTGTCC	**A034F**	TAGTTGTTTGGAGAAATTGG	**A046R**	TACTACCCTCACCCCAGC
A015F	TCTCACAGCCATTTCCTC	**A034R**	GGCAGGCAAACATTATAG	**A047F**	CAGAATGGTATGGGTTGG
A015R	CGTCATTAACCCACCATC	**A035aF**	TTGGAGCACTACGGCCTG	**A047R**	AAGGTTGAATCCAGGTGC
A016F	GGCCTATTTCTTTCTTTATG	**A035aR**	TTTGCCATTCTACCTCATGC	**A48F**	TGGGACAGATAGGCTCATAC
A016R	TACCATCAGTCTTAGGCG	**A035F**	ATCATTCAACCAGCTCTC	**A48F2**	GGACCTCAGCTTGCACAAAG
A017F	AGCTTTCATTGACTGGGC	**A035R**	ATGTGACCCTTTAGTATATG	**A48F3**	TGCTATACATGCCGAAAGTG
A017R	GGCCCTCTAGCAAATTCC	**A036F**	TTTTAGAGGAAGAGCACC	**A48F4**	TTCCCCCAATCATTGTGAAT
A018F	AAGGGTGGTATTTTCGGAC	**A036R**	ACCATTCCTTCACTTTCC	**A48R**	CTGAAGGAAACCAAATGTG
A018R	TCGAGACCGTAGTCTGCTG	**A037F**	TTATTTGGGATACCATGC	**A48R2**	CAATCGCTTGACTAGATGCC
A019F	CAAAATTCACACGGGAGG	**A037R**	TCTTTGATGGTAATGATCC	**A48R3**	ACCAAAGAGTCCTATGCCAG
A019R	AGAAGACCCCAGGATCAG	**A38E2F**	ATGCCAGCTGCATGAGAA	**A48R4**	CTTTGTGCAAGCTGAGGTCC
A020F	AGGAGCCATTGAGGGAGG	**A38E6R**	CACTGTTCCACTCCTTAG	**A48R6**	CAGGATGACAATGGGTACAAG

Table 36 Primers used for RT-PCR (contd)

Primer name	Sequence (5' ? 3')	Primer name	Sequence (5' ? 3')	Primer name	Sequence (5' ? 3')
A48R7	TGGGTACAAGAGACCAAATAGAAG	**A091R**	GGAGAAGCAGGGAGTCGG	**A129E2R**	TATAACTGCCCATGCACTT
A049F	TACACACTGCCCTATCCC	**A092F**	CAGTGGTAGGCAAGTGTG	**A129E3F**	GTTGTTCTGATGGGTGAC
A049R	TGTGTATGTGTGCGTTTG	**A092R**	TCCAACACTGACCTCTGC	**A129E3R**	TTTTCTTTTCTCCCTAAAGTC
A050F	AGACGAACGATTTGGATG	**A093F**	GAAATCATCACACTTGGGC	**A129F**	TCTGAGCCTAGAGGATACC
A050R	TTTGTTGAGTTTTGCTGG	**A093R**	AGATCCAATGCCCAATATG	**A129F2**	AAGAAACCTGGAGCCTG
A051F	GCTCAGGGCTTCAGGCAG	**A094F**	GGTGAAGGATGAGGGGTG	**A129F3**	TGATCTCCAATCTTCACAGC
A051R	GACAGAATCAATGGCCCAC	**A094R**	TTTGACCAGGGAACCACG	**A129F4**	AATTAAACAGAAGAGCACCC
A052F	GACCCATTCAAACCACCG	**A095F**	GTTTTTGGCACACAGAAGG	**A129F5**	AAGAAATATGGAGGTGAGC
A052R	TGCCTGCTGTCCCTTCTC	**A095R**	CCTCCTTTGATTGGTGAC	**A129F6**	GCCACACTCCTTCGCCAAC
A053F	TGCATTTGCCATTCTTCC	**A096F**	TAGGTAGGCATTATTGTCC	**A129F7**	GTTATTAAGATGAGTACACCTGCAG TCAATA
A053R	CATAGCAGGGAGGTTGGG	**A096R**	AAACCTTCATCTACTCCG	**A129R**	GATCTCAGAGGCAGGTTG
A054F	AATACAAGCAGAGAGCGG	**A097F**	AGTGGCTTTGGTGAGGGC	**A129R2**	ATAACCCCTGGCTCCCTG
A054R	GTTTCGGTTTTAGCAGGAG	**A097R**	TACCAAAGCAGCCCAAGG	**A129R3**	CCTCCACCCTGTCCCAAC
A055F	CATTTCCTGTTCCACTTCC	**A098F**	ATCGGTTGAATGTTTGGC	**A129R4**	AGCTTGAAGTGGCTAAAGTC
A055R	AGTGGGGGATGTTAGGAAG	**A098R**	CATGCCAGGSSSSGAAATG	**A129R5**	TGCTGTGAAGATTGGAGATC
A056F	GAATCCTACCTACACGGG	**A099F**	ATCCCTGGCTTCTAAACC	**A129R6**	AAATTGCCGTGATGATCC
A061F	TTCCTCTTCGCATCCTTC	**A099R**	TCATGGGTTGGCAAACTC	**A129R7**	GTGCTGGCAGGTCATGGAG
A061R	ACAAGAAAGCCATCACGC	**A100F**	ACCTACCAAATTGCAGCC	**A130F**	TTCCTTGACCTCCTTAGC
A062F	TGAAGGCATGAACGTGAG	**A100R**	CAAGAAGGCTGTAACTGG	**A130R**	AAAGTCCAGGTAACATTCC
A062R	CCAAATGGTGAAAACAGATC	**A101F**	TTACTTGTGCGTGAACCC	**A131F**	GCACTGTTGTTCGGGTCG
A063F	CTCTCCCGAGACATGGGC	**A101R**	TGTGTCCTTAGAAGTGCC	**A131R**	CACTCCCCATCACACTCC
A063R	GGAAGCCCTGACCCAAGC	**A102F**	TATGGACGCCCTAATCTG	**A132F**	ACAGAGGTCACATCAAGC
A064F	GGCCAGCCTCTAGTTTTG	**A102R**	TAAAGCCTATCAAAAGCCC	**A132R**	TTTCTCGTCCCGTGTTGC
A064R	CGAGTTTTCAGCCAGAGG	**A103F**	TTTGCTCACCTCCTTTCC	**A133F**	TGTCCTTGGTAATACTTCC
A065F	TCATTCAGACTTCCAGGC	**A103R**	AGAGGGCTGGAAATGTGC	**A133R**	GATAGTTGATGTCCCCAG
A065R	ATTACGGGGTCCTAGTGG	**A104F**	GTCCACTTTCAACCCCAG	**A134F**	GGATAAACCAGAAGATAGG
A066F	AGCAACACCACACCCTCC	**A104R**	AAAGGCAGCAAACGGTGG	**A134R**	TTAGCCCTAGTAACTTTGC
A066R	GCCCGATGTCAATTTCAG	**A105F**	GCTGTCATAAAGGGGGTG	**A135F**	GAATGATGGAGGAGGGAG
A067F	TGGATCTGTGGGCACTTG	**A105R**	ACTGTAGAGAGGGGCAAG	**A135R**	AGGGAAAGGTTACAGTGG
A067R	CCCACTGTCAACGCCTAG	**A106F**	ACTATTTGGGGGAGAGGG	**A136F**	GACCCCTGTGCCCTGCTC
A068F	GACCTGTGATGTTCGTGG	**A106R**	CTGGGTGAAGGTGGGGTC	**A136R**	GGTGCGTAGCCCAACAG
A068R	CCAGGAATGAAGCAACAG	**A107F**	ATTAGAAAACCAGACTCCTC	**A137F**	TGTTGAGTCTGAGGCTGG
A069F	TGGTGTGTTTTAGGTGGG	**A107R**	ATTGTCCCTTCCCATTATG	**A137R**	GCCCCTTGGTCCTGTTTC
A069R	CAACTAGAAAGCTGTGGG	**A108F**	GGGGATACATGGATCTTG	**A138F**	TCTTGTGTTGGGTGGGAC
A070F	AAAGGTCACACGGCATTG	**A108R**	TTAAACCTGAGCTTTAGGAC	**A138R**	ACTCAAATCGTAAGCATCC
A070R	TGGGGTCAGCAAAGGATG	**A109F**	TGGGCTGTTGAAATGACC	**A139F**	CTTCCCAGACCTACCAAC
A071F	TGTGCCAGGAAAGGAAGG	**A109R**	CAAAAGAAGGATGTATGCG	**A139R**	AGGTCTGGAAAGTCTCAC
A071R	TAGTCAGCAGCATCGGGG	**A110F**	ATTTTCCCCAAGCACAGC	**A140F**	ACATTTAGGAGGAGAGGG
A072F	AACAGAATAGCCAAGGTGC	**A110R**	TGGACTAGAAATGCAGGG	**A140F**	ACATTTAGGAGGAGAGGG
A072R	TCCTTTTCCTCTTGTGACC	**A111F**	ACCTTGAGTGGCGTGTGC	**A140R**	GTCTGTAGTCTGGGGATG
A073F	AAAGCAGGAGGATGTAGC	**A111R**	TCCCACCACTTTGCTCAG	**A140R**	GTCTGTAGTCTGGGGATG
A073R	CAATTAGGTGACATGAATGC	**A112F**	TTAGAAGGCATGGAATGG	**A143F**	TACTGTGGAGGTTATGTGAC
A074F	CTGAATCGCAATGGTCTG	**A112R**	CAGGGATCTTCAAGGACC	**A143R**	ATGTCCTTTGTGCTTCC
A074R	GGCGAGGCTTATTTCAGG	**A113F**	TTCTTCCCTGGAGCTTTG	**A144F**	TAATTCACAGAGCCCACC
A075F	GTCTTCTCCAACCCACAG	**A113R**	AGAAAGCAGGAGCACAATC	**A144R**	AATTGTGCCACTGTGCTG
A075R	GGGAGTCTTAGGTGGAGAG	**A114F**	CCCTCGCCTGCTTCCTTG	**A145F**	ATCTGGGCAAGGGCTGAG
A076F	AAACAGGGAAGTGAGAGC	**A114R**	CCCACGAGCCCAGAGACG	**A145R**	CGCTATATTCTGGTTGGGC
A076R	GAAGCCGAAATGAATGAGC	**A115F**	AGACAGTGGCAAAGACCC	**A146F**	CCTCAACACAAAAGTCTG
A077F	CCATGGTTATAAATGATAGC	**A115R**	GTTAGGCAATGGAGGTGG	**A146R**	GACATCATTTTCAGAGCC
A077R	TGAACAGCCATCCAGCAC	**A116F**	CTCTCAGGAAGGGCATC	**A147F**	CTTCCTTCTAGCCCATTCCAAT
A078F	GCCCCTATCCTTCCACAG	**A116R**	GTGGCGGAAATGTAGGTC	**A147R**	AGCCTGAGATTATACTTGAC
A078R	CGGTCCTTGGAGATCATG	**A117F**	GGATTTGATGCCAGGAAG	**A148F**	TAACCTTCCCCACAAACC
A079F	GGAGGAATAATGACCCAG	**A117R**	AGGCTTGCTTATTGCTGC	**A148R**	TCCTGGCCTCAGAAATACTG
A079R	ATTCAGTAAGGGGAGGTC	**A118F**	GTCAAACAGCGATGGGTG	**A149F**	GGAGGTGATTTTGGTGGG
A080F	TGCCAGGATTGTAGGATG	**A118R**	TGGGAGAATAGCCTGTGG	**A149R**	AAGCAGCCACTATCCATC
A080R	GTGTGCCCTCTGTAAATAG	**A119F**	GAGCCCCACTGACAAATG	**A150F**	CATTAACAAAGCAGTGGC
A081F	AATCTCCTGGTCTTTGGC	**A119R**	AGTGCAAGAGTGCCCATC	**A150R**	AAGCCAGAAACAAGAGCC
A081R	GTAGGTCCCCATTCTTCC	**A120F**	GACTGGACAGAGCGAATC	**A151F**	TCATAGGGTTAAGATGGC
A082F	CACCAAGAGGCTGAGATG	**A120R**	TTTCTCAAGCCTCTCCAG	**A151R**	CTGGTTGGGAACACAATTC
A082R	CCAATAGCACCTTTACCAG	**A121F**	CTCACTCTCCGCTTTTGC	**A152F**	CAGCTTGGATTTGAAAGG
A083F	TCCGACTCTAATTCAGGGG	**A121R**	CGGTCATTTTCCCAGTTTG	**A152R**	AGCACCCCTCTAACTTTCCC
A083R	AAATTCCAGGAAGGGCAG	**A122F**	ATGAAGCCGATGGGTGAC	**A153F**	AGTGAAGGTGACAAGGGC
A084F	TTCTGCTTGGCATGGAGG	**A122R**	ACCTCCACTCAAGACAGC	**A153R**	CCTCAATACCTCACACAAAC
A084R	TTCCTCTGCCATCCTTGC	**A123F**	ACCTGATGTTGCTGTTGG	**A154F**	AGTGAACATTTTCTGCATCT
A085F	ACAGGGGCTTAGAGATGC	**A123R**	TGGTGCGACACAATCCTC	**A154R**	TGCATACAGCAGGTAGATTG
A085R	AAAACACAGTCACGGAGG	**A124F**	CTTCTGGCCCCTCTCTTC	**A155F**	TCCTTGCCTGATGAGTTC
A086F	AGCTGTAACCCGAAATACC	**A124R**	TGGGATAGTTGGGAGAGG	**A155R**	ATAGTGTCCTTAAAATTGGCC
A086R	AAGTTTCAGATTCCCATTG	**A125F**	GCAATTCTCCAAACCAGG	**A161F**	TAAAGGTGGCTGCATGAG
A087F	CCAGCCTTTGCCGAATAC	**A126F**	AAGCAAGCCAGTCCTCAC	**A161R**	CGAGTGGGCATGAGATGG
A087R	TGAAAAGACATTGCCCCC	**A126R**	GAGGCAAAGTTCATCAGG	**A162F**	CCTGAGTACAGCATTGTG
A088F	CCTTCCCCCATATTACCC	**A127F**	TTCAGAACTTGGAAGCTGG	**A162R**	TACCTTGGCTGACCGTTC
A088R	ATGGCCTCAGGTATATGC	**A127R**	CGAGATTACTTCCCCATAC	**A163F**	CAGAACAAGTCAAAGGTACAC
A089F	TCGCAACCTGGACAAACC	**A128F**	CTCACATCCTTCTCAGCC	**A163R**	TTTCTATCCCCATTCTATTTC
A089R	ATGGCCTCAGGTATATGC	**A128R**	GTGGAATGTCAGGGAAATC	**A164F**	AGGAAGGGCCGGTGAAAG
A090F	ATAGGAGACAGACGAGTGC	**A129E1F**	ACATTGGGCTGCACATAG	**A164R**	CAAGTTTAAGGGCCACGG
A090R	CACTTGGCTGTTCCTATG	**A129E1R**	ATGCTACTCACATTAGAGAGATT	**A165F**	GCCTGTTCTGTGTATAGC
A091F	CCTTCTCGCTCTCCAGTG	**A129E2F**	TTAAGTTAATTGGGGGTTTA	**A165R**	CTGAGGGCTTGGTAATTC

Table 36 Primers used for RT-PCR (contd)

Primer name	Sequence (5' ? 3')	Primer name	Sequence (5' ? 3')	Primer name	Sequence (5' ? 3')
A166F	TTGAGGCATTTGAACAGGA	**A196R**	CATGAGTCCTTGTGCTAG	**B15R3**	GGCATCCGTAGTCAGTCCCTC
A166F2	ACTGCTGCCTCCCATGAC	**A197F**	CTAATTGCCGAGCACACC	**B15R4**	CAGATCCCTAGCATCACC
A166F3	CCGTAGGAGTCAGCGAAGGC	**A197R**	GCCACCAAGGATGATTTCC	**B15R5**	CCATGTCCCACCAGGCGGC
A166F4	CAGAGGGGATGGAGTAAG	**A198F**	TTTGAAATTCTGCCCTGCC	**B016F**	ACGTTTGATGAAAGAAGGTCG
A166F5	GCTGATGGGAGGATGTTGAG	**A198R**	TATGGCTCAACAGGGCACCC	**B016R**	GGACAGCACGÀGAAAATTCAAC
A166F6	ACCAGCAGTTCCACTGAGTC	**A199F**	TACCCAGCAGCAGGCCACCC	**B017F**	CGTGTCAATACCCTCTCAATG
A166F7	CTAGTTAGATGGGCACAGAG	**A199R**	CCGCTGCTCCTCGTCTCGCC	**B017R**	AATGCCTCTGGTTTTAGTTCC
A166F8	AATCCACCTCAGCAAGACCC	**A200F**	GCCACAAAGGGAAGCAGCGCC	**B018F**	CCTCCCCACCACGCTTCAC
A166F9	TCCGAGTTGCTGTGGTAGGG	**A200R**	TAGGTATTTGGACATGGGGCC	**B018R**	TGCAATTCACCAGCTACCAAG
A166F mus	CGACAGCATCTCAGATTTGG	**A201F**	GGAAAATGATTCCGGTTG	**B019F**	TCCAAGAGACTGATGCGGCTG
		A201R	GGCATGAAGTCAGCCTAG	**B019R**	TCTGTCCCAAAGTAGCGGTCG
A166R	TTGCCCACTTCATCTTTC	**A202F**	GTTGAGGGAAAGCTGAGGGCC	**B020F**	GAATCTCACCCTGTCGTCATG
A166R2	GATCTGCTGGTGCCAGTG	**A202R**	CCCAGGATGATTACTACCGCC	**B020R**	CCTCTCCCTCCCTCTTCAGTGC
A166R3	CAGCCAGAAGGATGAGAG	**A203R2**	CAGCTACGGCGTGGTTTA	**B021F**	ACCTTTACCAGCATCGCAGC
A166R4	GCTTCGCTGCCACTTCTC	**A204F**	AGTGCAGCATCAAAGAGGGCC	**B021R**	TGGACTGGCTGTTCGGACTTG
A166R5	GGCCTTCGCTGACTCCTAC	**A204R**	ACAGCAGCATCCACTTACGGCC	**B022F**	CGCAGAATGAACCACAAGAGC
A167F	ACAAACCAACAGGAGTATCG	**A205F**	GGAAGCCCATAGGAACCCGGCC	**B022R**	CGCAGAATGAACCACAAGAGC
A167F2	AGAGGAGTGAAAGTGGTC	**A205R**	GAATGAAGGTGCCCCAGGGCC	**B023F**	GCTGTACTGCAGCGCGTCG
A167R	TTAAATGGCTCCCTGCTG	**A206F**	CATGTCACATTTTGGGCC	**B023R**	GAAAACATACCATTCTCTCAAAGC
A168F	GTCAACAGCACCTCGTAG	**A206F2**	GGACCCCCTGTATCTGAG	**B023R2**	GGCTCCCTCAGTCTTAATCAC
A168R	GGCATGTTATTTCATTGG	**A206F3**	GAGTCTCTCCCTGTTGCC	**B024F**	CACCTCGCTGGGACTGATG
A169F	AGACACCTCCTCAATCGG	**A206R**	ATCTTTCTGGGCTACTGC	**B024R**	TGAAAACTATAAGGCAGCTTCC
A169R	CATGTGGATGGTGCAACC	**A206R2**	ATAGGCAAATAAGATATTCAC	**B025F**	CGGAGTGCAAGTGAGCTTCTC
A170F	AACCAACCCAACATTTCC	**A207F**	CATCACCTCGCAGAAGT	**B025R**	GGAGGCATGTGTTGCAGTGTC
A170R	GGATCTTTAGGCTTTCTG	**A207R**	ACACATGGAAGCTGCCAG	**B026F**	AGCCCATCCCATCCTTTATTG
A171F	TTGTAAGGTGCAGGCAGG	**A208F**	CTGCTCTCCACACACTGC	**B026R**	AAGCTCCTCTTCCTCATCCTC
A171R	GGGCTATAAAAGGGATGC	**A208R**	GAAGGCCCAGGACCTATC	**B027F**	ATGACCGTGCCTATGTGATG
A172F	ATGAATTAAGGATCAGGC	**A209F**	TGACCAAAGAAGCAGGAG	**B027R**	AAGTTAAACCCATTGATCCAG
A172R	CTTATAAATTTGGGGGAG	**A209R**	ACCTTCACTTCAGCCACC	**B028F**	GATGGAAAGGAAGTTGATGAAGG
A173F	ACAATCGTCAATGGAAGG	**A210F**	CACACCATCCTAACACCG	**B028R**	GAACCAACAGGAAATGATGCG
A173R	ATCCCTTTATCCTGGCTC	**A210R**	AACTGGTAAAGAGCAAGG	**B029F**	CTGCGAAGCCTCCTCCTCAAG
A174F	AAAGAAAGTGCTAACAGGAC	**A211F**	AAGCAGACACATGCCCAG	**B029R**	TTTATTGGGTGTTGGGGATTC
A174R	TTCTGCAGTGCTAATGAGC	**A211R**	TTTAAGCCATCAGGAGAC	**B030F**	GAGTGATAACAGTGCCGAGTG
A175F	AGTGATGTCTTCGGAGGG	**A212F**	GTCTAGTTCCCAAATTCC	**B030R**	TTTCCTCATGGCTCGGATCTG
A175R	ACTGGATGGAGAGGAGGC	**A212R**	CCCAAGAGTCACATAGCA	**B031F**	GTGACCAAAGCAGTGAAGAAG
A176F	TCCTCAGAACTACTGCAAAG	**A213F**	AGCCTCCAGCCAAGTGTG	**B031R**	GTCCCCAGGCTCTACCACAAC
A176R	CTCAAGTGGATTTCAGCAG	**A213R**	ACACCGCGTCCCCAGGAG	**B032F**	AGTTAATATGGGCGTCCAAGC
A177F	TCTCCCCAAAATGACACGG	**A214F**	CGGAGCGGAGTAAGCGTC	**B032R**	GCCCGCCATCCCAGGTCTC
A177R	TGTGGTAGATCATTTTGTGTTG	**A214R**	TAATGCTTGAATTGGATG	**B033F**	GGATGTTGGATGTGTGGCTGC
A178F	GTACACCCTATACATGCCG	**A215F**	TTGTCTTCACTGGCCTTC	**B033R**	GTGATCCAAGCCTCGGTTTTC
A178R	CAAAAAGTGCTGTTAGCGG	**A215R**	TAATGCTTGAATTGGATG	**B034F**	TGAGAAACCTGCCCTTGACTG
A179F	GGAAGTTTGTAAGCACGC	**A216F**	ATAACCTTGGTCATGCAG	**B034R**	TCTTTGATAGTCCGAATGTGC
A179R	AGCTGAGGGATGGCCGAG	**A216R**	TTACATGATTCCAAGAGGG	**B035F**	CCGGAGATTTAAGTGGGCTA
A180F	GCTTGGCGCTGCTGTGAGCC	**A217F**	CCTCTGCGTGTGGATGTC	**B035R**	ACACAGAGAGTGAACGCAATG
A180R	GGACTACAGCGAGCGGAGGCC	**A217R**	GAATTTCTCTCCTTGGCC	**B036F**	ATATGCCAAGCGGCAATATAG
A181F	CCACTTGATGCTCCTTGCGCC	**A218F**	AGCTAAGGCAAAAGGCAG	**B036R**	GGTGGAGGGCAATGGGTTCAG
A181R	GGGCCATATCAACTTCACGCC	**A218R**	ATGTTGGGTCGGCTTCTG	**B037F**	GGCTCTCAGGCTTTTGTAATC
A182F	ATGTTCAGCAATGGGGACGCC	**B001F**	CGGCACCACAGTAGTCCAAAG	**B037R**	TGTTGTGTCTTGCCTGTTTCC
A182R	CAGGCACACAGTAAGCACGCC	**B001R**	CTTCCAGCTTCCGTATACTG	**B038F**	CATGCCTTGTGGTTCATTCAG
A183F	CCTCATCATTCCTGGTGGGCC	**B002F**	TAAATGAGGTGAGGTGCTTGGT	**B038R**	ACCTGTTCCTTTCCTGTCCTG
A183R	CATCAGAAGCACGCACAGCC	**B002R**	GGTGGAGGGAGTGCGATAG	**B039F**	TCCAGTCATCTTTGAGGTAGC
A184F	TTGTCCCATTTCAGAGGCGCC	**B003F**	TCCACAAGCACCAAGAGTCG	**B039R**	GAGTATGTTTGTCAGCCTTGC
A184R	GGAGGTGCCACGTTATGGGCC	**B003R**	GTTTGCCAGCCTCCTTTCTC	**B040F**	GAGGGTAGGTTGGCAGTATT
A185F	ACCGAAACTGAGGACAAG	**B004F**	TGATGAGAAACCAGGAAAACG	**B040R**	ATAGTGGGTGATGGGAGAGAC
A185R	TGCAGGGAGAAGTTATCC	**B004R**	GCCACAGGGGCTTCACTT	**B041F**	TCTTGCATTTCTGAAGCTCTG
A185R2	CTCCTTAATTGCCACAGC	**B005F**	CAGCTTTCTCAGCAACTGTCC	**B041R**	CAGCACAAGGGAATAGCAGAC
A186F	TAACTAAACTGGCTCCCC	**B005R**	TGACATCCCTAAGAGCCAAGT	**B042F**	GGATCGCGTAGAGGAACCTTC
A186R	AAAGAGGTGAAGGGGAGC	**B006F**	GGACATAACTGCCTACCACACA	**B042R**	GTGATATTTCTTCCTCGTGCC
A187F	GATAAGCCTGGTTTCTGG	**B006R**	GACTGTGTTGGGTTTCTTGAG	**B043F**	TCTCCCACCACTGACTTCCTC
A187R	GTTTTTTCCTAGCCACCC	**B007F**	GGCACCACATTTCACCTCAAG	**B043R**	CTAGGGGTGGTGATGGACCAG
A188F	GGTTTCTCTCTCCAGCAG	**B007R**	GGAAGACAAATCAGAGGACGG	**B044F**	CAGAGAGGGGACCAGGAAAG
A188R	AAAGTGAATGCTGGGTGG	**B008F**	AGTAGACCGCAGATGTGAGC	**B044R**	TTTAACCATACCTTGTGCTCG
A189F	AGGTATGCTGATTGGTGC	**B008R**	GGGATCATGGTAGCTTATGTG	**B045F**	AGCGGATGTATGTTCGGGTTG
A189R	TAGTCTCATTCACCTGCC	**B010F**	CATGAAGTGAAAAGCGGGAAG	**B045R**	GGCTTGGAGAGTGGTGGGTTG
A190F	TTAGCAGTTTGTGGTGGG	**B010R**	CAAGCACTCTGAACTGTCCTG	**B046F**	TCCTAGGTAGCCCAATTCCAG
A190R	ACAGGCAGAAGACAAAATG	**B011F**	CCGCAGATTGGCATTCAGTG	**B046R**	ACATCTTGCTGCCTTCGGGAG
A191F	TAAAGAGCAAGGACTGGC	**B011R**	CAGCTCTTAGGGTGAGGGCAG	**B047F**	GAAGCATGTTCACCATTTATTTTC
A191R	ATGCTGAATCTCCTTTTGC	**B012F**	CGATTGACTTTGCTTTCCCTA	**B047R**	TTTTAAAGGCACCTAAATAGAGC
A192F	TAGAGTGGGAGAAACAGGGCC	**B012R**	CCCAACTCTCCCTCCATTCAC	**B048F**	GTGTCTTTCTTTGCTCCGTCC
A192R	TTCCGACACTTCATCAGCGCC	**B013F**	CCACGTAACAAAATTCTGAGG	**B048F**	GTGTCTTTCTTTGCTCCGTCC
A193F	GCAAAGGGTAAATTACATGGCC	**B013R**	TCTGCTCAGCTCGCTTTGG	**B048R**	TTCTGCTGGCTGATTTGGTGG
A193R	AACACATTTCCGAGAGGCGCC	**B014F**	GGGTGAGGTGCTGTTGATGGT	**B049F**	ATCTGTTCTTTGAGTGTGTGG
A194F	TTTAGGAGAAGCATTTGTGGCC	**B014R**	CTAGCACTTTTGAGCATTTG	**B049R**	GAGAAAGGATAAACCAAGCAG
A194R	TTGTGAGGGAAGGTAAGG	**B15F**	GGAAGGACTGTTGCTGGTATG	**B050F**	TAACAGCCAGATTCCATTTTG
A195F	GACATCCTAAAACAACATC	**B15F2**	CCGCCTGGTGGGACATGG	**B050R**	CATGAGGGCAGTAACACTAAG
A195R	AATTAACCCATGTCAGCC	**B15R**	CTAGCACTTTTGCAGCATTTG	**B051F**	CTCCAGACCTGCTCCATAG
A196F	GTTAAAATGTGGCACTGTG	**B15R2**	GTCTGATGAGGGCTACTTTGTG	**B051R**	AAGATAGATTTTGAGACCTGC

Table 36 Primers used for RT-PCR (contd)

Primer name	Sequence (5' ? 3')	Primer name	Sequence (5' ? 3')	Primer name	Sequence (5' ? 3')
B052F	CCTGTTTGGGCCAATTAAGAC	**B072R**	TTCCCAAGGACCGTAGATGC	**L35R**	CAAACACTATCCGAAGCCAG
B052R	AGAAGTGTTGTGATCTGGCTTG	**B073F**	CTGGCTTGAGTGTCTGCATTG	**L36F**	TTAGCACATACATCCACTTG
B053F	ATTCTGCTCCCCCACACATCC	**B073R**	CAGCTCTCCTTTCCTCCTTCC	**L37F**	AACTGGCTACTCTTTCAACG
B053R	GTTGCAGTGGCTGTGATTGTC	**B074F**	CGGAGAGGGGTGTGTGAGTG	**L37R**	CTTCATCCTCATTGTCTTCG
B054F	TTAGAGAACAGCCAGCACAGG	**B074R**	CCAGGGACCGTGACACACAG	**L38F**	TCCACAACGATGCTTTCTA
B054R	GAGGCGATGTCAGGGAGG	**L02F**	AATGAGCGTGTAGAAGAAAG	**L38R**	AACAAACCTTCACACAAATC
B055F	AAGTGTGGAGAGTTAGGGCAG	**L03F1**	CCTGCCTGCTATTGGTCAC	**L39F**	TAACACCCCAGCCATTGATT
B055R	CTGCAAAGGCCAATCACTCAC	**L03R**	TGAACTCTGGGTCTTTGTAG	**L39F2**	AAGAAAAGGAGATTAGGGTACAG
B056F	GATAGAGGTGATAGGAAATGTTG	**L05F**	GGTTCAGAATTTTCACAGCT	**L39F3**	TGTCATGCTAGGAAAGTCGAAC
B056R	AGCAAATATAGAAATTGAGGTTG	**L05R**	CTGGTGTTTGTGTCCTATTC	**L39F4**	CAAGGTAGGTGGGAGAAGGTG
B057F	TGGAGAAACTTGTGTGTATGC	**L10 f**	TTCCTTCCCTTCAGGTACTG	**L39R**	CAGGAGGAGCACAAGCAAG
B057R	TGCTTCAATCTTATCTCCCAC	**L10 r**	TCCACATTGATATAGGCCAC	**L39R2**	GGAAGTTTATTAGGAAGAATTGG
B058F	AAGCTTGACAGCAAGACAAAC	**L14F**	TGCTGTTTGTAAGGCTGGAG	**L39R3**	CAATACAAGATCCGCATAGCAG
B058R	TTGACAAATTTCAGGTTTTGC	**L14R**	TCTCTTTGATGACACGGCTC	**L39rR4**	GCACCTGGCCACCAAGAG
B059F	GTCCTCCCACAAATGCTCCG	**L16F**	TGAGGAAGAGGAAGCAGAAC	**L40F**	AAGAACAAAAAAGAGGATGC
B059R	GGGCCTGAGAGGGGTCACC	**L16R**	CTGAACCTATCCATCCCAC	**L40R2**	AAGGCAGCCACTACAGCTAG
B060F	CAGTCGTTGGAAGCAGAAATG	**L17**	GATAGAAGAATAGGGGCAAG	**L47F**	CAACTTCAGAGACAACCCAC
B060R	CTCGATTTCATCTAGGCAACC	**L17R**	GAATGGACAACACAAATAGG	**L47R**	CCGTATTTATGGAGATTGGT
B061R	ACCAAAATCTTAGTCCAGCAGT	**L18F**	GTAGAAGAGTTGAAGGCTGC	**L48F**	CCATGAAGAAGATGACATTG
B062F	AGCCCCTCTATCTCTTCCTTC	**L18R**	CATACAAGGAGGAAGACTGG	**L48R**	AGACAGGATCTCTGTGCTAG
B062R	TGAAGGCTCAATCTGGTGGTC	**L20F**	CTGATGGAAACCGCTGAAG	**L50F**	GAGATGAAGGTGAAGGTGTC
B063F	TACCACAGGGGAACACAGAAG	**L20R**	CTTCAATGGCTTCCTCTGC	**L50R**	CATAGAAAGGTGAAGTGCTG
B063R	TGCGTGTGCTTGGCTGGTG	**L21**	TTTACTTGCTGTGGTTGTGG	**L54F**	TTATCTGTCCTCCAAAGTGA
B064F	CGATCAAAGAAGAGGGACAGG	**L21R**	TCTGAGGCATAGCAAAACAC	**L54R**	TGATTGTCGTTGCTGCTTTC
B064R	CAGGCCACTGAAATTCAAATC	**L23F**	AGTTCTCCTGATGTTGTTCC	**L56F**	GATTCTTCACAGGCTCAGG
B065F	GCCGCCTAATGTCTTTTGTTC	**L23R**	GCCCACACATCATAAGGTAC	**L56R**	ACCTGGCTAACTCATCCTTC
B065R	CTGGTCCCTTTGGTCTCTATG	**L24F**	GTTGTGGGTGACTCTTTTGG	**L63F**	CCAGAAGCTCCAACATTGTC
B066F	GGGGGGTTCTTCCACTTTAG	**L24R**	GTGAGGTCTGTTCCAATAGC	**L63R**	CTAACAAGACAATGAGGCAG
B066R	GTCGCCTCGGTCTTCCTTG	**L25F**	GAAAAGACCCAGAAGGATG	**L72F**	TCCGTGGTGATGAAGAGTTG
B067F	CCTTATTCACACCTTCCTTCC	**L25R**	AGCCATCTAACAGGTCATC	**L72R**	TCCCAGGTTCTCACTCTTTC
B067R	CCAAGCCAGATGAAGTCCTCC	**L27F**	CAGGCAAGAAGGTGTTAGTG	**L78F**	TAAGCTCCCTAACTGCCTTC
B068F	CCCGATTGAGTTTCTGCCTTC	**L27R**	GGTCTGGAATGAATGGAAAG	**L78R**	AGAAACATCATCCAGGGTCG
B068R	ATGCAGGACAGAGTACCTCAGC	**L28R**	GCCCTACAGAATGAATACAC	**L86F**	GTTCAGAGCGAAGGAGGAG
B069F	AAGCTAAACCAAGAGCACCAG	**L28R**	TATCATTCACAAGTTTCCCA	**L86R**	GCTCCAGTCTCCTATCATCC
B069R	GTCCTGCAAACTGAAGCAAAG	**L30F1**	GGCTGCTACCATTAAACAAC	**L88F**	TATACCACTGGCACAACAAC
B070F	GTGTGTTTGCTCAGGGCTCAG	**L30R**	CAGTGAGAAGTGCCATAAGC	**L88R**	ATGATGATGTCCCAACTCCT
B070R	GAGAGGTGTAATGAAGGGTCC	**L32F**	GTGATCCAGAAAGTTGATGG	**L92F**	GCTTACTGAGGACTTCTTTC
B071F	TGAAACCAAGACCCAAGAGG	**L32R**	CATTACCAAGACAACCCCTC	**L92R**	TGTTCTCCCTTTCTGTCCTG
B071R	GATCTTGGATACCTTCCTGCC	**L33F**	CTTCCACTTCTTCCATCTGC	**L93F**	TGAATGACGCACTGGAGAAG
B072F	CTGCAGGCAACACCAGATCC	**L33R**	TGCTGGTTTCACAATACTGC	**L93R**	ATTTACCAAGCAAGTGCAGC
B072R	TTCCCAAGGACCGTAGATGC	**L35F**	ACCTTTGCCTTCTCTTGCTC		

Table 37 Primers used for qRT-PCR

Gene amplified	Primer name	Sequence (5' ? 3')	Size of product	Final Mg^{2+} cc	Annealing temp.	Real-time measure	PCR efficiency
CAMTA1, KIAA0833	A004F2	AGCCTCTAACCCCTTCTTC	243	3	57	72	100%
	A004rR3	AGGGTCGGCCCTTGTATT					
ERO1LB	A017rF2	GCTTCCAGAGAATAGTCC	98	3	61	72	93%
	A017rR2	CTCTTATACTTGTAGAAAGCCTTCC					
PSMD11	A059rF2	ACGAAGCTGCTCTGGAAA	97	3	57	72	85%
	A059rR2	CCGCTACAGATCCAACTCTATG					
KIAA1796	A084rF2	GGCTGAAGTGATTGCAGG	97	3	57	72	87%
	A084rR4	ATTCCCATGATGTTCTCGAA					
BC042097	A085rF4	GCAGAGGCAAAGAATGGA	86	3	57	72	79%
	A085rR3	ACAGAGTAGATGAGGAAAGGG					
FLJ13993	A106rF3	GGTATTTGTCTCGGAAGTGG	88	2	60	86	89%
	A106rR5	GTGGAAGAGAAGATGGA					
KIAA1263, DKFZp566D234	A109rF2	AAGCAATTTGACAGAGTGGAT	178	3	57	72	104%
	A109rR2	ATGACTGAGGAACGCACTTAT					
A111, BC016878	A111rF3	GGCTTCGCATAATTTTCAT	86	3	62	82	84%
	A111rR4	AACGCGTTCAACACACACA					
C12orf3	A126rF6	TGTCATAATCTTCCAGAACC	146	3	57	72	83%
	A126rR8	GCTTCTCATGCTCCTCTTTCA					
AK054981, DKFZp547H074	A150rF2	TGAAACATTTCAGTTCATTCTC	147	3	57	72	85%
	A150rR2	GCTTCCCTTTTCTGCTATGTTCT					
CHD1	A165rF2	CAAGCAAGACAGCAGATATTAC	79	3	57	72	93%
	A165rR2	TGACCTGTGATCTCTACTCC					
C1orf32	A166rF10	TGATGGAAAGACTGGCGGAA	85	3	64	80	95%
	A166rR7	GGATGAGAGGCAGATATAACAA					
ORC2L	A168rF2	AGGCATTCCTCGTCAATAG	101	3	57	72	91%
	A168rR2	CTCTACTCCATCAGTTCCCTTC					
AA057097, in intron of	A169rF2	ACACTCTACTGAGATCCTCC	133	3	63	82	96%
	A169rR2	CAGCAGATGTTGCTGCACA					

Table 37 Primers used for qRT-PCR (contd)

Gene amplified	Primer name	Sequence (5' ? 3')	Size of product	Final Mg^{2+} cc	Annealing temp.	Real-time measure	PCR efficiency
DKFZp761D221	A177rF2	TCAACAATGTGCAGTTCC	91	3	57	72	93%
	A177rR2	TGTTGTTCAGCATTCCAGA					
STK35 alt splice, AL844428	A203rF7	CCTGATGCCTTTGAACTTGA	104	3	64	80	94%
	A203rR3	CCGAGGAATCGACCTAGTTT					
BC035234	A205rF2	ACTTTGCCCGTGAGATGA	99	3	61	81	92%
	A205rR2	CACTGCATCCTTGACCTTGAT					
AK056484	A206rF5	TTCTCCAGCTACATTTAGGGAA	85	2	64	73	92%
	A206rR3	GTTGGTGGGATTTTGGGTCATT					
SF3B3	A211rF2	TCTGAACATCCCCCTCTCT	84	3	57	72	95%
	A211rR2	GTCTCCATCAATCACATTCTTCAC					
CRYPTIC	A213rF3	TCAATTTGGGAAACAGCTATC	66	2	63	74	
	A213rR3	CAACCTTGGTGACTTCCTCT					
FLJ31564	B001F5	TGGTTGAGAAGGGACAGGATT	70	2	67	78	95%
	B001R4	AAGTCTTGGTTTGTTCTGATGAG					
BC034603, B15	B15R4	TGCCAGATCCCTAGCATCACC	114	3	61	81	84%
	B15rF4	CGCAACCTGCGTCAAAG					
AK091467, B030	B030rF3	GGTCAGCTACCTCCTTCA	79	3	57	72	87%
	B030rR5	TTCGTCTTCAGCCACCTTA					
CLASP2, KIAA0627	L02rF2	CTGGCAGTAAAATGAAGCTA	158	3	57	72	95%
	L02rR2	GATGAGGGTGGTCTATCTGT					
KIAA1380, AL831917	L05rF3	GAGCACTTCATCAGGTTCA	113	3	57	72	93%
	L05rR3	CAAAAGTCTCAGTTCCTGTGT					
C14orf129	L10rF1	GCATCATCAAGAATGGAAA	92	3	61	79	95%
	L10rR1	CCTTCAAAACCGTTCAGCTC					
KIAA1576	L11rF4	GTGTGGACATCGTTTTGGATT	110	3	61	81	95%
	L11rR2	GTTACCATGTTGGATGAGCCATA					
FLJ33282, BC029611,	L14rF3	AATGGTAGCCTCTGGGAA	80	3	57	72	97%
	L14rR2	AGAAGAGCCACCACGAGAT					
PLCD4	L16F2	GCGTTTTGTGGTAAATGGA	99	3	58	78	93%
	L16R2	ATGTGGCGGTAACCTTGTTG					
SSX2IP	L17rF2	CTGCATATCTGAACATAGTTCAA	128	3	64	72	91%
	L17rR1	CTTCCTGTGATCCAGGTCTT					
FLJ13305, AK023367	L18rF1	GCAGACATCGAAGCAGATGA	186	3	61	72	95%
	L18rR1	CATCCTTTCCTTTTCGCTCTT					
EKI1	L21F3	AAGTTACTGAAAAGGAGGTAGAAA	75	0	65	75	91%
	L21R3	CCCAAAAGAAATGAGAAGCCAA					
C20orf103	L23rF1	GTCCACATCCAACCTTT	67	3	57	72	102%
	L23rR1	CCACTGGGCATTTATGCTCTTC					
SLC1A2	L24rF2	GTCACCATGCTCCTCATTCT	125	3	66	86	89%
	L24rR2	CCAAAAGAGTCACCCACAAC					
ABCC5 splice variant	L25rF4	TCCAAAGGAAGGCTGAAC	118	3	60	80	87%
	L25rR4	GTAGTGGAACCAGTGACAGAAG					
SV2B	L28rF2	CTGATTTACCTCGTCAGCTTC	118	3	61	83	95%
	L28rR2	AGATTAGCATGGAGCCACCAA					
L33, BC029061	L33rF2	TTGCATGGAAGATCGTGAAC	109	3	64	80	94%
	L33rR2	CATACTTCTGCTGCGTCCAA					
L35	L35F1	CTGGCTTCGGATAGTGTTTG	162	3	61	80	86%
	L35rR2	GAGAATCCCATAATGCAGA					
DKFZp434C0631	L36rF3	CCCTTGTTTTCAAGCTCTTC	92	3	57	72	100%
	L36rR2	CGCAATAAATCCAGTATGGTGGT					
L37, BM668448	L37rF1	CAAAGGATGCACAGCAAC	116	3	57	72	87%
	L37rR1	TCGGGTGACAGAAGCAA					
L38	L38F	TCCACAACGATGCTTTCTA	221	3	58	72	84%
	L38R	AACAAACCTTCACACAAATC					
L39	L39F	TAACACCCCAGCCATTGATT	171	3	61	80	95%
	L39rR4	GCACCTGGCCACCAAGAG					
ZPBP	L40F2	CACATGAAAGGTCAAGAAGC	258	3	61	81	86%
	L40R2	AAGGCAGCCACTACAGCTAG					
MGC14816	L47F	CAACTTCAGAGACAACCCAC	148	2	64	80	89%
	L47rR1	GTTGACTTGGTGATCTGCTCT					
L48	L48rF1	CAAGTGGAGAAACAGGAAG	83	3	57	72	97%
	L48rR1	CTAGAATGGCAACCAGGTTCA					
L52, BC040189	L52rF1	ACACAGTTTTCTGCTCCCT	95	3	62	84	92%
	L52R	AGAAGCCCAAGAGTCCCTG					
L54 long splice variant	L54rF2	CCTTCCTTTCCTCTGTGGATT	149	3	59	77	98%
	L54rR2	CGGTTCTTTTACTGGATCATA					
L56	L56rF1	TTGATCCTTGTCTGGCTCTGT	148	3	61	72	95%
	L56R	ACCTGGCTAACTCATCCTTC					
DKFZp547C176	L63rF1	CTTTCAGCAAACATTCTGCTA	97	3	57	72	100%
	L63rR1	CTCCCTCACACGATTAAATTCC					
H2AV variant 2	L72F	TCCGTGGTGATGAAGAGTTG	154	3	57	72	97%
	L72rR1	TCAGCACACATCCCAGTAGAA					
FLJ30499	L78F	TAAGCTCCCTAACTGCCTTC	175	1	56	72	100%
	L78R	AGAAACATCATCCAGGGTCG					
KIAA1579	L88rF1	GGGAGAACCACCAAAAGA	99	3	57	72	93%
	L88rR1	GGTTTTACTTGCAGGGGATGA					
L93, AF086541	L93F2	CACTGGAGAAGCTCAAC	118	3	62	82	98%
	L93R2	CGAGGCTGCTGAATAATGTAG					

Table 37 Primers used for qRT-PCR (contd)

Gene amplified	Primer name	Sequence (5' ? 3')	Size of product	Final Mg^{2+} cc	Annealing temp.	Real-time measure	PCR efficiency
Housekeeping genes							
ACTB	ACTB F	GACATCCGCAAAGACCTGTA	136	2	65	72	96%
	ACTB R	CAGGAGGAGCAATGATCTTGA					
B2M	B2M F	TAAGCAGCATCATGGAGGTTT	70	2	66	72	97%
	B2M R	AGCAAGCAAGCAGAATTTGGA					
TBP	TBP F1	GGTTTGCTGCGGTAATCAT	106	2	66	83	100%
	TBP R1	CTGGACTGTTCTTCACTCTTGG					
RPL13A	RPL13A F1	AAGCCTACAAGAAAGTTTGC	116	3	65	84	98%
	RPL13A R1	TAGTGGATCTTGGCTTTCTC					
HPRT1	HPRT F	CACTGGCAAAACAATGCA	94	3	65	82	95%
	HPRT R	GGTCCTTTTCACCAGCAAGCT					
GUS	GUSB6	GATCCACCTCTGATGTTCAC	160	3	65	83	97%
	GUSR	TATTCCCCAGCACTCTCGTC					
SDHA	SDHA F	TGGGAACAAGAGGGCATCTG	86	2	64	78	91%
	SDHA R	CCACCACTGCATCAAATTCATG					

2. Abbreviations

5'-RACE	rapid amplification of cDNA
A	adenine
aa	Amino acid
AAP	abridged anchor primer
acc.no.	accession number
AD	autosomal dominant
AMD	Age-related macula
ApoE	apolipoprotein E
APS	ammonium persulphate
AR	autosomal recessive
AS	Aminosäure
AUAP	abridged universal
BLAST	Basic Local Alignment Search
BLAT	Basic Local Alignment Tool
bp	Base pair
BSA	bovine serum albumine
C	cytosine
cDNA	coding DNA
CIP	calf intestinal phosphatase
cM	Centi Morgan
CNS	central nervous system
Cxorfx	chromosome x ORF x
CYMD	dominant cystoid macular
DAPL1	death associated protein-like 1
dd	double destilled
dd H_2O	double-distilled water
DEPC	diethylpyrocarbonate
DMSO	dymethylsulfoxid
DNA	deoxyribonucleic acid
DNase	deosyribonuclease
dNTP	Desoxynucleosidtriphosphate
ds	double strand
DTT	dithiothreitol
E	efficiency
EDTA	ethylenediaminetetraacetic
ER	endoplasmic reticulum
EST	Expressed sequence tag
G	guanine
g	gram
GRM7	metabotropic glutamate
hrs	hours
Hs.	*homo sapiens*
htgs	unfinished high throughput
INF-?	interferon gamma
kb	kilo base pairs
kDa	kilo Dalton
LB	Luria-Bertani
LD	long-distance
LISCH7	Homo sapiens liver-specific
LTR	long terminal repeat
M	molar
µl	microliter
µM	micro molar
min	minutes
ml	mililiter
mM	mili molar
mm	millimeter
MMLV	Moloney Murine Leukemia
MOPS	4-Morpholinepropanesulfonic
mRNA	messenger RNA
nM	nano molar
NMD	nonsense-mediated mRNA
no.	number
OLB	oligo labeling buffer
OMIM	Online Mendenlian Inheritance
ORF	open reading frame
PAA	polyacrylamid
PCR	polymerase chain reaction
pfu	plaque forming units
pH	
qRT-PCR	quantitative PCR
retNEIBank	retina cDNA collection from
retSSH	retina suppression subtracted
RFRP	RFamide-related peptide
RLM-RACE	RNA Ligase Mediated Rapid
RNA	ribonucleic acid
RP	retinitis pigmentosa
RPE	retinal pigment epithelium
RT-PCR	reverse transcriptase PCR
RZPD	German Resource Center for
SAGE	Serial Analysis of Gene
SDS	sodium dodecyl sulfate
sec	seconds
SNP	single nucleotide
ss	single strand
SSC	sodium-saline citrate
SSCP	Single Strand Conformation
SSPE	sodium phosphate buffer
T	thymine
T_a	annealing temperature
TAP	tobacco acid pyrophosphatase
TBE	tris-borate
TEMED	N,N,N ,N -
THC	Tentative Human Consensus
T_m	melting temperature
Tris	Tris-
U	units
UTR	untranslated region
UV	ultraviolet
V	volt
VLDL	Very Low Density Lipoprotein
VN	virtual Northern
vol	volumes
W	watt

IX List of publications

Thesis-related publications

Stöhr H, Mah N, Schulz HL, Gehrig A, Frohlich S, Weber BH. 2000. EST mining of the UniGene dataset to identify retina-specific genes. Cytogenet Cell Genet 91:267-277

Mah N, Stöhr H, Schulz HL, White K, Weber BH. 2001. Identification of a novel retina-specific gene located in a subtelomeric region with polymorphic distribution among multiple human chromosomes. Biochim Biophys Acta 1522:167-174

Schulz HL, Stöhr H, Weber BH. 2002. Characterization of three novel isoforms of the metabotrobic glutamate receptor 7 (GRM7). Neurosci Lett 326:37-40

Schulz HL, Stöhr H, White K, van Driel MA, Hoyng CB, Cremers F, Weber BH. 2002. Genomic structure and assessment of the retinally expressed RFamide-related peptide gene in dominant cystoid macular dystrophy. Mol Vis 8:67-71

Stöhr H, Schulz HL, Stojic J, Fröhlich S, Berger C, Weber BHF. 2002. Towards generation of a gene anatomy atlas of the human retina in health and disease. Progress Report 1999-2002 German Human Genome Project 100-101

Poster presentations at meetings

Schulz HL, Stöhr H, Mah N, Weber BHF. 2000. Cloning and characterization of two novel retina-specific genes. German Human Genome Project Meeting 2000, Heidelberg, Germany

Stöhr H, Mah N, Schulz HL, Fröhlich S, Weber BHF. 2000. EST mining of the UniGene dataset identifies genes specific to the human retina. German Human Genome Project Meeting 2000, Heidelberg, Germany

Schulz HL, van Driel M, Cremers FPM, Stöhr H, Weber BHF. 2001. Cloning of a retinal gene encoding RFamide-related peptides: a candidate for cystoid macular degeneration on chromosome 7p. 10th International Congress of Human Genetics, Vienna, Austria

Stöhr H, Schulz HL, Mah N, Weber BHF. 2001. Identification of candidate genes predisposing to age-related macular degeneration by systematic EST based expression profiling. 10th International Congress of Human Genetics, Vienna, Austria

Schulz HL, Stöhr H, Fröhlich S, Berger C, Stojic J, Weber BHF. 2002. Identification of gene preferentially expressed in the human retina using an expressed sequence tag (EST) approach. 13. Jahrestagung der Deutschen und Österreichischen Gesellschaft für Humangenetik, Leipzig, Germany

Stojic J, Gehrig A, Schulz HL, Wagner M, Weber BHF. 2002. Identifying novel retina-specific genes by characterising a suppression subtracted cDNA library highly enriched for retinal genes. 13. Jahrestagung der Deutschen und Österreichischen Gesellschaft für Humangenetik, Leipzig, Germany

Schulz HL, Stöhr H, Stojic J, Fröhlich S, Berger C, Weber BHF. 2002. Towards a characterization of the human retinal transcriptome. 1st Symposia of the National Genome Research Net, Berlin, Germany

Schulz HL, Stöhr H, Stojic J, Weber BHF. 2003. Towards comprehensive characterization of the human retinal transcriptome. 53rd Annual Meeting of the American Society of Human Genetics, Los Angeles, USA

X CURRICULUM VITAE

PERSONAL INFORMATION

Name:	Heidi Schulz
Date of birth:	February 18th, 1973
Place of birth:	Geneva, Switzerland
Nationality:	German
Marital status:	Married

SCHOOL EDUCATION

1979 - 1983	Elementary school (1st - 4th grade), Uruguay
1983 - 1984	Elementary school (5th - 6th grade), USA
1984 - 1986	Junior high school (7th - 8th grade), USA
1986 - 1989	High school, Argentina

UNIVERSITY EDUCATION

1991 - 1996	Biochemistry study at the University of Buenos Aires, Argentina
1999 - 2000	Diplom thesis: 'Identification of ten novel genes expressed predominantly in human retina by candidate gene selection from gene indexing projects'. Institute of Human Genetics, University of Wuerzburg, Germany
2000 - 2003	Doctoral thesis: 'Towards a Comprehensive Description of the Retinal Transcriptome: Identification and Characterization of Differentially Expressed Genes'. Institute of Human Genetics, University of Wuerzburg, Germany

PROFESSIONAL EXPERIENCE

1995 – 1996	Hospital laboratory residence in Clinical Chemistry, Bacteriology, Blood extraction and Hematology, River Plate Hospital, Argentina
1994 – 1997	Teaching assistant for the course 'Biological Chemistry', School of Pharmacy and Biochemistry, University of Buenos Aires, Argentina
1996 – 1997	Residency in the cytology department, River Plate Hospital, Argentina
1997 – 1999	Diagnosis of human papilloma virus infection using hybridization techniques and PCR. Pathology and Cytology Institute, Essen, Germany