Bioinformatic and molecular approaches for the analysis of the retinal pigment epithelium (RPE) transcriptome

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

vorgelegt von
Faisal Mohamed Fadl El Mola
aus dem
Sudan

Würzburg 2003

Eingereicht am: 13.08.2003

Mitglieder der Promotionskommission:

Vorsitzender: Prof. Dr. RAINER HEDRICH

Gutachter: Prof. Dr. BERNHARD WEBER

Gutachter: Prof. Dr. ERICH BUCHNER

Tag des Promotionskolloquiums: 24.09.2003

Doktorurkunde ausgehändigt: . . . .

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Diese Dissertation wurde weder in gleicher noch in ähnlicher Form zu einem anderen Prüfungsverfahren vorgelegt.

Es wurde zuvor kein anderer akademischer Grad erworben.

Würzburg, den 13.08.2003

Acknowledgments

This study has been carried out at the Institute of Human Genetics, University of Würzburg School of Medicine, Germany.

I wish to express my thanks to Prof. Bernhard Weber for his invaluable comments and guidance. I would like to thank Prof. Dr. Erich Buchner for taking the responsibility of being my Ph.D. supervisor. I am very grateful to Prof. Dr. med. Holger Höhn, Director of the Institute of Human Genetics, for his helpful discussions and advise.

I am grateful to Andrea Gehrig for constructing the RPE cDNA library, Birgit Halliger-Keller and Andrea Rivera for the technical assistance. I am thankful to Faisal Rahman for providing Northern blot for the 2 clones that were characterized in this thesis. I must also thank Nicole Mehr and Claudia Berger for help in sequencing the DNA clones. Thanks and appreciation goes to the staff and graduate students of the Institute of Human Genetics for being friendly, as well as cooperative. I am very thankful to Mrs Neumann for her help handling the administrative aspects during the course of the study. Finally, I would like to express my deepest gratitude to my family for their patient and encouragement.

TABLE OF CONTENTS

ABBREVIATIONS

ZUSAMMENFASSUNG

SUMMARY 1

I INTRODUCTION ………………………………………………… 3

1.1 Structure and function of the retinal pigment epithelium (RPE) …… 3

1.2 The RPE and hereditary diseases……………………………………… 6

1.3 The RPE and age-related macular degeneration (AMD) ………………... 7

1.3.1 Evidence for genetic and environmental factos in AMD susceptibility…. 8

1.3.2 Oxidative stress and AMD pathogenesis ………………………………... 10

1.3.3 Growth factors and AMD ………………………………………………. 10

1.3.4 Role of apoptosis in AMD ………………………………………………. 11

1.3.5 Vitamin A metabolism and AMD ………………………………………. 11

2. Genetic approaches to correlate the genotype with phenotype ……… 12

3. Computational methods for identifying genes ………………………... 14

3.1 Public databases and sequences analysis tools …………………………... 18

3.1.1 Sequences analysis resources …………………………………………… 19

3.1.1.1 BLAST ………………………………………………………………….. 19

3.1.1.2 dbEST …………………………………………………………………… 19

3.1.1.3 OMIM …………………………………………………………………... 20

3.1.1.4 dbSNP …………………………………………………………………… 20

3.1.1.5 Pfam …………………………………………………………………….. 20

3.1.2 Database Retrieval Tools ……………………………………………….... 20

3.1.2.1 LocusLink ..……………………………………………………………... 20

3.1.2.2 Source …………………………………………………………………… 21

3.2 Genome Analysis and annotation …………………………………………... 21

4. SNPs as genetic markers ……………………………………………….. 22

4.1 SNPs and association studies …………………………………………… 22

4.2 Haplotype blocks and tag SNPs ………………………………………….. 23

4.3 Identification of New Polymorphisms …………………………………… 23

II. OUTLINE OF THE PRESENT RESEARCH …………….. 25

III. MATERIALS AND METHODS ……………………………. 26

1. cDNA library construction …………………………………………….. 26

2. EST amplification and sequencing ……………………………………. 27

3. Bioinformatics ……………………………………………………………27

3.1 Design and Implementation of a Relational Database Management System (DBMS) ………………………………………………………… 27

3.1.1 Structured Systems Analysis and Design Method (SSADM) …………... 28

3.1.2 Logical Data Modelling (LDM) ………………………………………… 28

3.1.3 Relational Data Analysis (RDA) ……………………………………….. 30

3.1.4 Physical Database Construction (MS Access) …………………………...32

3.2 BLAST searches ………………………………………………………… 33

4. Identification and characterization of genes …………………………... 34

4.1 RNA extraction and first strand cDNA synthesis ………………………. 34

4.2 Polymerase Chain Reaction (PCR) ……………………………………… 35

4.2.1 PCR_x Enhancer solution ……………………………………………………. 36

4.3 Reverse Transcriptase (RT) PCR ………………………………………... 36

4.4 Northern blotting…………………………………………………….. 37

4.5 Cloning of PCR products ………………………………………………… 37

4.5.1 Ligation of RPE01-D2 ……………………………………………………… 38

4.5.2 Transformation of the ligated RPE01-D2 ………………………………. 38

4.5.3 Clones amplification and sequencing …………………………………… 39

5. SNP genotyping for RDH12 (RPE01-D2) …………………………….. 39

5.1 SNPs identification ……………………………………………………… 39

5.2 Determination of allele frequency of the identified SNPs ..………………40

IV. RESULTS ……………………………………………………… 42

1. RDBMS design and construction ……………………………………… 42

2. Analysis of the SSRPE cDNA library…………………………………... 59

2.1 Phase I: Analysis of 1002 ESTs …………………………………………… 59

2.2 Phase II: Analysis of additional 1377 ESTs …………………………….. 61

3. Functional profiling of the reported known RPE genes……………… 63

4. Comparison to similar studies 63
5. Cloning and characterization of RPE01-D2 and RPE6-C10 66
5.1 RPE01-D2 66
5.2 RPE06-C10 68
5.2.1 Genomic structure and expression pattern 68
5.2.2 Analysis of the predicted cDNA sequence 70
6. SNP genotyping for RDH12 71

V. DISCUSSION 74
1. Candidate gene approach and SNP analysis 74
2. RDBMS construction 75
3. SSRPE cDNA library analysis 77
4. Blast searches and genes prediction 80
5. From genes to functions 81
6. Comparison to similar studies 83
7 Cloning and characterization of RPE01-D2 and RPE6-C10 86
7.1 RPE01-D2 86
7.2 RPE06-C10 87
8. Generation of a SNP map of RPE01-D2 (*RDH12*) 88

VI. CONCLUSIONS AND FUTURE DIRECTIONS 90

VII. REFERENCES 91

VIII APPENDIX 103
1. ESTs representing known genes in the SSRPE cDNA library classified by function 103
2. Publication and Presenations 109
3. CURRICULUM VITAE 110

ABBREVIATIONS*

µl	microliter
10-D9R	RPE microtitre plate number 10, row D and column 9 reverse
A	adenine
ABI	Applied Biosystem
AD	Alzheimer's disease
AMD	Age-related macular degeneration
AREDS	Age-Related Eye Disease Study
Arg	Arginine
ATG	Start codon
Aβ	β-Amyloid
B. taurus	Bos taurus
BLAST	Basic Local Alignment Search Tool
bp	base pair
C	cytosine
CAP3	Contig assembly program
cDNA	omplementary DNA
CDS	coding sequences
CH	choroid
CNV	choroidal neovascularization
COGs	Clusters of Orthologous Groups
DNA	deoxyribonuclease
dbEST	Database of expressed sequence tags
DBMSs	Database management systems
dbSNP	Database of Single Nucleotide Polymorphisms
DEPC	diethyl pyrocarbonate
DHPLC	denaturing high performance liquid chromatography
DHRD	Doyne honeycomb retinal dystrophy
dNTP	desoynucleosidtriphosphate
DOS	Disk operating system
DTT	eithiothreitol
E. Coli	Escherichia coli
ECM	extracellular matrix
EDCCS	Eye Disease Case Control Study
EDTA	ethylenediaminetetraacetic acid
EMBL	European Molecular Biology Laboratory
e-PCR	electronic PCR
ESTs	expressed sequence tags
F	forward primer
FNF	First Normal Form
G	guanine
g	gram
G3PDH	glyceraldehyde 3-phosphate dehydrogenase
GCL	ganglion cell layer

GEO	Gene Expression Omnibus
Gln	Glutamine
GSP	gene specific primer
GUSB	ß-glucuronidase
HMM	hidden markov model
HTG	high-throughput genomic
htSNPs	haplotype tag SNPs
ID	identifier
INL	inner nuclear layer
IPE	iris pigment epithelium
IPM	interphotoreceptor matrix
IPTG	Isopropyl-ß-D-thiogalactopyranoside
IS	inner segments
IVS	intervening sequence
kb	kilo base pairs
kDa	kilo dalton
LB	Luria-Bertani
LCA3	Leber's congenital amaurosis 3
LD	linkage disequilibrium
LDM	logical data modelling
LD-PCR	long distance polymerase chain reaction
Lib	leucine-rich repeat protein induced by beta-amyloid
LINE	long interspersed nuclear element
LOC	Locus
LRAT	lecithin retinol acyltransferase
LRR	leucine rich repeat
LRRC15	leucine rich repeat containing 15
LRRCT	leucine rich repeat C-terminal domain
min	minute
ML	Malattia Leventinese
ml	mililiter
MOPS	Morpholinopropanesulfonic acid
mRNA	messenger RNA
MS	Microsoft
NAD	nicotinamide adenine dinucleotide
NADP	nicotinamide adenine dinucleotide phosphate
NCBI	National Center for Biotechnology Information
NEI	National Eye Institute
NHGRI	National Human Genome Research Institute
^{o}C	Degree Celsius
OMIM	Online Mendelian Inheritance in Man
ONL	outer nuclear layer
ORF	open reading frame
OS	outer segments

PCR	Polymerase chain reaction
Pfam	Protein families database
POS	photoreceptor outer segments
PUFA	polyunsaturated fatty acids
QBF	Query by Form
R	reverse primer
RDA	Relational data analysis
RDBMS	Relational database management system
RDH11	retinol dehydrogenase 11 (all-trans and 9-cis)
RDH12	retinol dehydrogenase 12 (all-trans and 9-cis)
RDH5	retinol dehydrogenase 5 (11-cis and 9-cis)
RNA	ribonucleic acid
ROS	reactive oxygen species
RPE	Retinal pigment epithelium
RPE01-D2	RPE microtitre plate number 1, row D and column 2
RPE06-C10	RPE microtitre plate number 6, row C and column 10
RPE1	bovine retinal pigment
RPE24-D11	RPE microtitre plate number 24, row D and column 11
RS1	retinoschisis (X-linked, juvenile) 1
RT-PCR	reverse transcriptase PCR
SAGE	serial analysis of gene expression
SDR	short-chain dehydrogenases/reductases
SDS	sodium dodecyl sulphate
SMART	Switching Mechanism At the 5' end of RNA Transcript
SNF	Second Normal Form
SNP	single nucleotide polymphism
SQL	structured query language
SSADM	Structured Systems Analysis and Design Method
SSC	standard saline-citrate
SSCP	single-strand conformation polymorphism
SSH	suppression subtractive hybridization
SSPE	Sodium chloride sodium phosphate + Ethylenediaminetetraacetic acid
SSRPE	suppression subtractive RPE
T	thymine
TEAA	triethylamine acetate
TGA	stop codon
TNF	Third Normal Form
Tris	Tris-hydydroxy-methyl amino-methane
TUNEL	dUTP nick end-labeling
UCSC	University of California, Santa Cruz
UK	United kingdom
UNF	Un-normalised Form
USA	United States of America
UTR	untranslated regions
UV	Ultraviolet

VBA	Visual Basic for Applications
VEGFR-1	vascular endothelial growth factor receptor-1
X-Gal	5-bromo-4-chloro-3-indolyl-ß-D-galactopyranoside

---

*Please see appendix (VIII, 1) for gene symbols mentioned in this thesis.

Zusammenfassung

Es besteht ein grosses medizinisches Interesse an der Identifizierung von Genen, welche an der Entstehung komplexer, häufiger Krankheiten des Menschen beteiligt sind. Eine solche Krankheit ist die alters-korrelierte Makuladegeneration (AMD). Die AMD ist eine der häufigsten Ursachen für den Verlust der Sehfähigkeit im Alter von über 75 Jahren. Obwohl die Erblindung bei der AMD letztlich durch das Absterben von Photorezeptor-Zellen in der zentralen Retina bedingt wird, gibt es genügend Hinweise dafür, dass die Pathogenese der AMD ihren Ausgang vom retinalen Pigmentepithel (RPE) nimmt (Liang and Godley, 2003).

Ziel dieser Arbeit war die Identifizierung und Charakterisierung von RPE-spezifischen Genen als Beitrag zur umfassenden Charakterisierung des RPE-Transkriptoms. Darüberhinaus war es Ziel der Arbeit, die mögliche Rolle der RPE-spezifischen Gene bei der Entstehung der AMD zu explorieren. Ausgangspunkt der Arbeit war eine RPE-spezifische, bovine cDNA Bibliothek, welche in der Arbeitsgruppe auf der Grundlage der SSH-Technik (Diatchenko et al, 1996, 1999) hergestellt worden war. Die SSH-Technik gestattet die Anreicherung von differentiell exprimierten Genen bei gleichzeitiger Normalisierung redundanter Sequenzen. Mit Hilfe des Software-Programms CAP3 (Huang and Madan, 1999) wurden insgesamt 2379 ESTs gruppiert und geordnet. 1,2% der 2379 RPE-ESTs enthielten Vektor Sequenzen und wurden daher von der weiteren Analyse ausgeschlossen. 5% der RPE-ESTs wiesen Homologien zu multiplen Chromosomen auf und wurden daher ebenfalls von der weiteren Analyse ausgeschlossen. Die übrigen 2245 ESTs wurden in 175 Contigs und 509 Singletons gruppiert, woraus sich Hinweise auf insgesamt 684 putative Einzelgene ergaben. 343 dieser 684 Klone zeigten jedoch keine Homologien zu humanen orthologen Sequenzen. Ursache für die fehlende Homologie muss in der grossen Zahl der Klone gesehen werden, bei welchen nur die 3´untranslatierten verglichen wurden. Im Gegensatz zu den kodierenden Sequenzabschnitten kommt es in den nicht-kodierenden Regionen in der Regel zu einer relativ raschen evolutionären Divergenz und damit zum Verlust der Homologie (Sharma et al, 2002). Durch zusätzliche Sequenzierung und Sequenzvergleiche der kodierenden Bereiche dieser 343 Klone lassen sich möglicherweise weitere RPE-spezifische Gene finden.

Um die grosse Anzahl der im Rahmen des RPE-Projektes generierten Daten bearbeiten zu können wurde eine sehr effiziente und Benutzer-freundliche Datenbank auf Grundlage des RDBMS-Moduls etabliert. Dieses System gestattet die interaktive Bearbeitung der

gespeicherten Daten im Query-Format. Darüberhinaus können die Daten in beliebiger Weise annotiert und verbunden werden.

Nach Abzug der 343 nicht-homologen cDNA Klone von den 684 putativen Einzelsequenzen verblieben 341 Kandidaten-Sequenzen. 2 dieser Sequenzen wurden als putative neue RPE-spezifische Gene einer weiteren Analyse zugeführt. Dabei wurde zunächst die RPE- bzw. Retina-Spezifität dieser Kandidaten-Sequenzen mit Hilfe der RT-PCR Analyse bestätigt. Als Basis für zukünftige Fall-Kontroll- und Assoziationsstudien wurde eine SNP-Genotypisierung eines dieser zwei Klone (ursprüngliche Bezeichnung: RPE01-D2; derzeitige Bezeichnung: RDH12) durchgeführt. Die direkte Sequenzanalyse umfasste 23.4 kb und ergab insgesamt 12 SNPs, von denen sich 5 als hoch-informativ erwiesen. Auf dieser Grundlage können zukünftig Allel-Frequenzen zwischen Kontrollpersonen und AMD-Patienten ermittelt und verglichen werden. Zukünftig werden darüberhinaus real-time PCR Methoden zur Expressionsanalyse der verbliebenen Kandidaten-Klone eingesetzt.

Zusammenfassend liefert die vorliegende Arbeit einen Beitrag zum Verständnis der genetischen Grundlagen der RPE-Funktionen und trägt zur Aufklärung der Rolle von RPE-spezifischen Genen bei der Disposition zur AMD bei. Zusätzlich ergaben sich Hinweise auf Kandidatengene, welche möglicherweise in der Pathogenese der AMD eine Rolle spielen.

SUMMARY

There is substantial interest in the identification of genes underlying susceptibility to complex human diseases because of the potential utility of such genes in disease prediction and therapy. The complex age-related macular degeneration (AMD) is a prevalent cause of legal blindness in industrialized countries and predominantly affects the elderly population over 75 years of age. Although vision loss in AMD results from photoreceptor cell death in the central retina, the initial pathogenesis likely involves processes in the retinal pigment epithelium (RPE) (Liang and Godley, 2003).

The goal of the current study was to identify and characterize genes specifically or abundantly expressed in the RPE in order to determine more comprehensively the transcriptome of the RPE. In addition, our aim was to assess the role of these genes in AMD pathogenesis. Towards this end, a bovine cDNA library enriched for RPE transcripts was constructed in-house using a PCR-based suppression subtractive hybridization (SSH) technique (Diatchenko et al., 1996, 1999), which normalizes for sequence abundance and achieves high enrichment for differentially expressed genes. CAP3 (Huang and Madan, 1999) was used to assemble the high quality sequences of all the 2379 ESTs into clusters or singletons. 1.2% of the 2379 RPE-ESTs contains vector sequences and was excluded from further analysis. 5% of the RPE-ESTs showed homology to multipe chromosomes and were not included in further assembly process. The rest of the ESTs (2245) were assembled into 175 contigs and 509 singletons, which revealed approximately 684 unique genes in the dataset. Out of the 684, 343 bovine RPE transcripts did not align to their human orthologues. A large fraction of clones were shown to include a considerable 3´untranslated regions of the gene that are not conserved between bovine and human. It is the coding regions that can be conserved between bovine and human and not the 3’ UTR (Sharma et al., 2002). Therefore, more sequencing from the cDNA library with reclustering of those 343 ESTs together with continuous blasting might reveal their human orthologoues.

To handle the large volume of data that the RPE cDNA library project has generated a highly efficient and user-friendly RDBMS was designed. Using RDBMS data storage can be managed efficiently and flexibly. The RDBMS allows displaying the results in query-based form and report format with additional annotations, links and search functions.

Out of the 341 known and predicted genes identified in this study, 2 were further analyzed. The RPE or/and retina specificity of these two clones were further confirmed by RT-PCR analysis in adult human tissues. Construction of a single nucleotide polymphism (SNP) map was initiated as a first step in future case/control association studies. SNP genotyping was carried out for one of these two clones (RPE01-D2, now known as RDH12). 12 SNPs were identified from direct sequencing of the 23.4-kb region, of which 5 are of high frequency. In a next step, comparison of allele frequencies between AMD patients and healthy controls is required. Completion of the expression analysis for other predicted genes identified during this study is in progress using real time RT-PCR and will provide additional candidate genes for further analyses.

This study is expected to contribute to our understanding of the genetic basis of RPE function and to clarify the role of the RPE-expressed genes in the predisposition to AMD. It may also help reveal the mechanisms and pathways that are involved in the development of AMD or other retinal dystrophies.

I INTRODUCTION

1.1 Structure and function of the retinal pigment epithelium (RPE)

The retinal pigment epithelium (RPE) consists of a monolayer of cuboidal cells that is situated between the fenestrated vasculature of the choriocapillaris and the photoreceptor cells of the neurosensory retina (Marmor, 1998). The RPE is polarized with the apical RPE cell surface facing the photoreceptors, and its basal infoldings functionally linked to the choroid via Bruch's membrane, a five-layered structure of the extracellular matrix. The apical and basal membranes of the RPE cells are differentiated with regard to receptoral and ion channel properties. The sodium-potassium pump is present on the apical membrane, whereas the chlorid-bicarbonate exchange transporter is on the basal membrane (Figure 1) (Marmor, 1998).

The RPE gets its name from the *melanin* pigment that is present within cytoplasmic granules called melanosomes (Marmor, 1998). These granules begin to fuse with lysosomes and break down with aging process. This is the reason why the elderly fundus appears less pigmented. The melanin absorbs light that has been captured by the photoreceptors and prevents excessive light scattering within the eye. Although the molecular mechanism of photoprotection is not fully understood, melanin has been postulated to act as a cellular antioxidant (Sarna et al., 2003). New findings show that the melanin granules are connected to the lysosomal degradation pathway, and that a deficit of melanin pigment is associated with age-related macular degeneration (AMD).

Another RPE pigment is *lipofuscin* that accumulates in RPE cells with age, it can occupy up to 19% of cytoplamic volume by the age of 80 (Feeney-Burns et al., 1984). Lipofuscin is thought to be accummulated as a result of the highly oxidative retinal environment, which may oxidize proteins and polyunsaturated fatty acids (PUFA) that are abundant in photoreceptor outer segments (POS) and autophagocytic debris, rendering them refractory to lysosomal action (Shamsi and Boulton, 2001). This leads to progressive lipid accumulation seen with aging and the formation of basal laminar and basal linear deposits and drusen that are present in lesions of AMD (Curcio and Millican, 1999). It has been reported that basal linear and soft drusen are diffuse and focal deposits, respectively, of the same membranous material (Sarks et al., 1980) and that membranous debris in Bruch membrane (BM) is associated with AMD. Although, the role of basal deposits in the development of late AMD, characterized by choroidal neovascularization (CNV) and geographic atrophy of the RPE, remains controversial, the presence of basal linear deposits is thought to place an eye at risk for late AMD (Curcio and Millican, 1999). Up to 40 hydrolytic enzymes are present within lysosomes of which cathepsin D together with cathepsin S has been shown to be important in the breakdown of POS (Eldred, 1998).

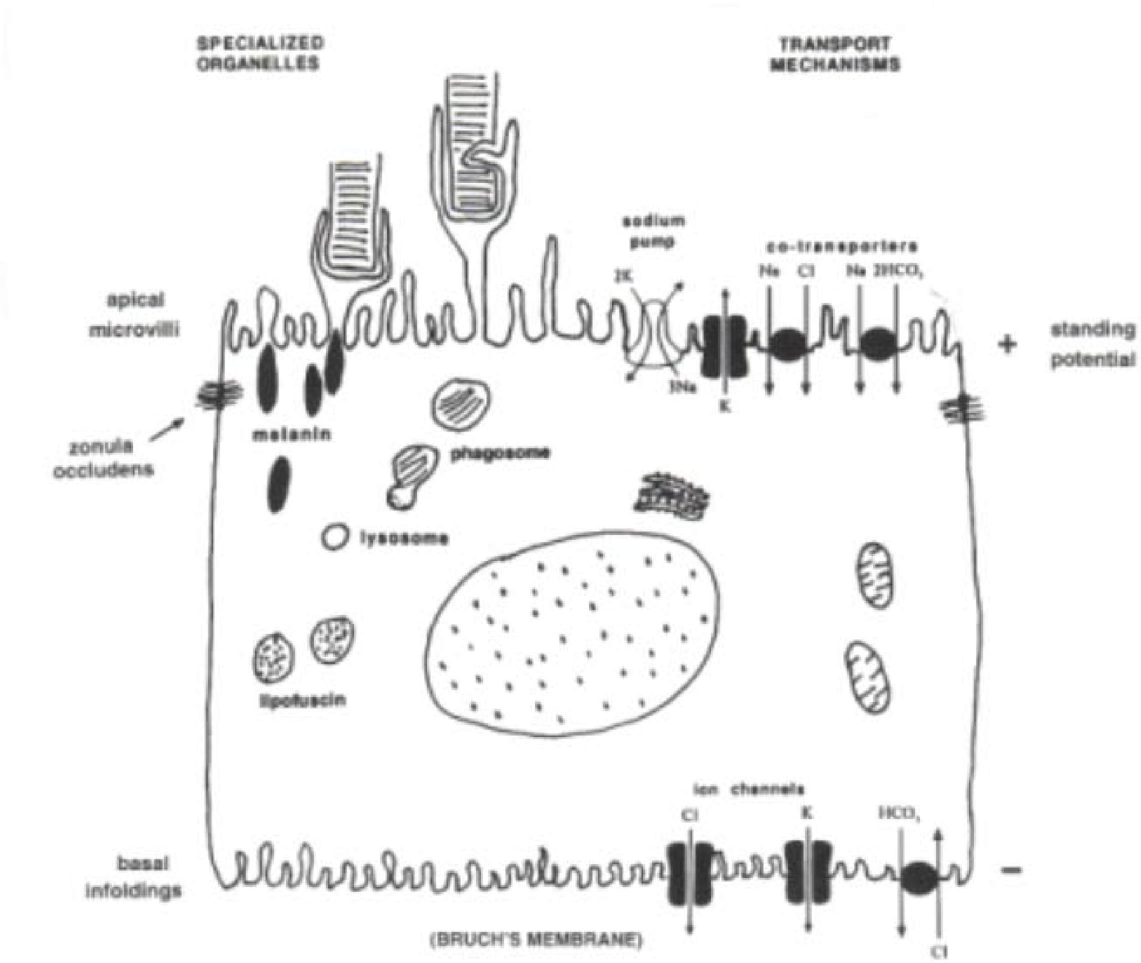


Figure 1 Diagram of a retinal pigment epithelial (RPE) cell. Note the microvilli projecting from the apical surface and the extensive infoldings on the basal surface of the RPE cell. Melanin granules are more numerous in the apical region, whereas lipofuscin particles are more abundant in the central and basal areas. Other organelles shown include those present in all cell types (nucleus, mitochondria, lysosomes, endoplasmic reticulum, and Golgi). In addition, the RPE contains specialized organelles such as melanin, phagosomes, and phagolysosomes (from Marmor, 1998).

The essential physiologic functions of the RPE include the selective diffusion and transport of ions, metabolites, and serum components to the outer retina (Hughes et al., 1998); transport, storage, and processing of vitamin A and its derivatives; the absorption of scattered light by RPE melanin granules; and the synthesis of basement membrane components, including fibronectin, laminin, collagen, and glycosaminoglycans, all of which are important to RPE adhesion to and maintenance of Bruch’s membrane. Table 1 summarizes some of the major known functional characteristics of the RPE (Marmor, 1998).

Table 1 Physiological roles of the RPE*

Pigment functions

Light adaptation and screening
Detoxification and binding
Lipofuscin accumulation
Antigenic properties

Environment and metabolic control

Blood-retina barrier
Transport of nutrients and ions
Dehydration of subretinal space
Synthesis of enzymes, growth factors, pigments
Interaction with endocrine, vascular and proliferative factors

Visual pigment cycle

Capture and storage of vitamin A
Isomerization of all-trans to 11-cis vitamin A

Interphotoreceptor matrix and retinal adhesion

Specialized matrix ensheathment of rods and cones
Metabolic control of adhesion

Outer segment phagocytosis and aging

Phagocytosis of outer segments tips
Digestion and recyling of membrane material
Aging effects: lipofuscin, drusen
Deposits and alterations in Bruch’s membrane

Electrical activity

Responses to light-induced ionic changes
Responses to light-induced chemical signals
Nonphotic responses to chemical agents

Repair and rectivity

Repair and regeneration
Immunologic interactions
Scarring and pigment migration
Modulation of fibrovascular proliferation

* This table was adopted from (Marmor, 1998).

On its apical side the RPE contributes to the formation and maintenance of the interphotoreceptor matrix (IPM) that is critical for retinal adhesion. Recent studies have shown that the IPM is a highly structured with chemically independent domains surrounding the rods and cones (Marmor, 1998). The matrix serves as a pathway for retinoids, nutrients and other substances to cross the subretinal space, and it appears to be an important domain for the elaboration of binding and receptoral proteins that assist control the photoreceptor environment.

1.2 The RPE and hereditary diseases

The RPE is involved in a wide variety of congenital, inherited, and metabolic disorders, although few have been identified in terms of specific cellular dysfunction at the level of the RPE. This might be due to the intimate interplay between retina and RPE pathophysiology, and overlap between the expression in the retina or RPE of genetic abnormalities. However, as the physiological functions of the RPE become well understood, we are realizing that the RPE is involved in the pathophysiology of many disorders that were once identified as affecting purely the choroid or the retina. Congenital abnormalities of the RPE include albinism and congenital hypertrophy of the RPE. Albinism is a subgroup of disorders arising from a reduction of melanin pigment in the RPE. Pigment deficits in the RPE also lead to a reduction in the number and distribution of photoreceptors and other retinal cells by affecting the mechanisms controlling cell proliferation during the early development of the retina. Another common condition involving RPE is retinitis pigmentosa, a group of disorders characterized by slow degeneration of the photoreceptors with hereditary transmission. RPE degeneration is present in many other retinal dystrophies such as Stargardt's disease, pattern dystrophies, choroideremia, and photic maculopathy. Retinal detachment resulting from loss of the adhesive function of the RPE can also cause photoreceptor degeneration.

A number of studies have attempted to transplant RPE to the subretinal space in various animal models (Lavail et al., 1992; He et al., 1993; Sheng et al., 1995). RPE transplants have also been attempted in humans with macular degeneration using cultured and freshly isolated homologous fetal RPE, the results have been disappointing in most parts due to immunological rejection (Algvere et al., 1999).

Transplantation of autologous RPE cells is not feasible because it is difficult to obtain autologous RPE cells, and the cells could possibly carry a genetic defect because they would be obtained from a patient bearing the disease to be treated. Recently, it has been suggested that autologous iris pigment epithelium (IPE) cells which can be easily obtained, may be an appropriate substitute for RPE cells for transplantation into the subretinal space to treat diseases in which the RPE loses its function (Thumann, 2001). In vitro studies have shown that IPE cells have the potential to carry out many functions characteristic of RPE cells, e.g., retinol metabolism (Thumann et al., 1999). Understanding the factors which control RPE growth and proliferation is central to effective RPE or IPE transplantation.

1.3 The RPE and age-related macular degeneration (AMD)

AMD is a prevalent cause of legal blindness in industrialized countries and predominantly affects the elderly population over 75 years of age. AMD is a degenerative condition of the cone-rich central retinal called the macula, and its prevalence is rising with increasingly aging populations (Leibowitz et al., 1980; Klein et al., 1992). Despite the high incidence and severity of vision impairment, only a limited percentage of AMD patients are amenable to treatment (Ciulla et al., 1998).

The clinical and histopathological features of AMD include a strong relationship with age, and the presence of pigmentary disturbances, drusen, thickening of Bruch's membrane, and basal laminar deposits with normal visual acuity to large areas of RPE atrophy or choroidal neovascularization. Drusen are subretinal pigment epithelial deposits (Figure 2) that are characteristic of but not uniquely associated with AMD (Hageman et al., 2001). Small, hard drusen with well-defined edges are present in more than 90% of eyes in old people and are usually not considered to be pathological. On the other hand, larger soft drusen are diagnosed as early AMD (Bird et al., 1995; Curcio and Millican, 1999). Late AMD is divided into two clinical subtypes: (i) atrophic or ‘dry’ AMD, characterized by accumulation of drusen within and under the RPE and atrophy of the macular retina and RPE; and (ii) exudative or ‘wet’ AMD, typified by invasion of abnormal blood vessels into the subretinal space from the choroid and subsequent disciform degeneration. Although the vision loss of AMD results from photoreceptor damage in the central retina, the initial pathogenesis involves degeneration of the RPE (Green et al., 1985; Spraul et al., 1996; Zarbin, 1998). While the causes of AMD are not conclusively known, it thought that multiple factors may be involved including environmental, nutritional and genetic components. Thus, AMD is considered a multifactorial disease caused by genetic as well as environmental factors.

1.3.1 Evidence for genetic and environmental factors in AMD susceptibility

Several studies suggest that there is a genetic component to AMD pathology. Twin studies have shown a higher concordance of AMD phenotypes among monozygotic (100%) than dizygotic (40%) twins (Meyers et al., 1995). Both clinic-based and population-based family studies have reported an increased risk of AMD among first-degree relatives of affected individuals (Klaver et al., 1998; Seddon et al., 1997). The prevalence of AMD among first-degree relatives of subjects with AMD, particularly with exudative disease, is greater than among first-degree relatives of subjects without the disease. The data suggest that macular degeneration has a familial component and that genetic or shared environmental factors, or both, contribute to its development (Seddon et al., 1997). Family linkage studies have identified one possible locus for AMD on chromosome 1q (Klein et al., 1998).

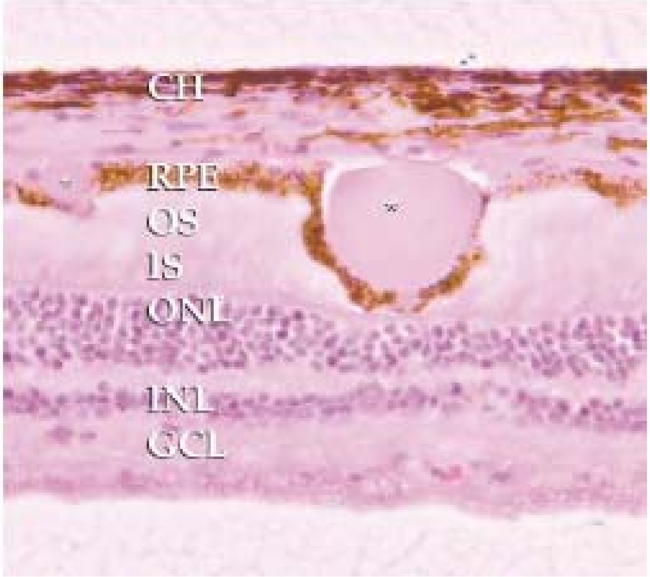


Figure 2. Light micrograph illustrating the appearance and location of "hard" drusen. Drusen are located in the sub-RPE space between the RPE basal lamina and the inner collagenous layer of Bruch's membrane (Asterisks); OS, photoreceptor outer segments; CH, choroid; IS, inner segments; ONL, outer nuclear layer; INL, inner nuclear layer; GCL, ganglion cell layer (from Hageman et al., 2001).

The association of AMD with genetic factors is further supported by identification of the disease-associated gene for Stargardt disease (ABCR or ABCR4), a rare early onset macular dystrophy, which shows some overlapping clinical and pathological features with AMD (Allikmets et al., 1997; Briggs et al., 2001). The significance of ABCR gene sequence changes in AMD has been challenged by a number of studies that either identified methodologic deficiencies in the design of the study (Dryja et al., 1998; Klaver et al., 1998) or failed to reproduce a significant association between ABCA4 and AMD (Rivera et al., 2000; Webster et al., 2001). The contribution of ABCA4 and its extent to AMD pathogenesis still remains to be demonstrated. Several other genes have been associated with inherited retinal dystrophies that also share some clinical manifestations and features with AMD. These include the gene encoding the tissue inhibitor of metalloproteinase-3 (TIMP3) resulting in Sorsby fundus dystrophy (Weber et al., 1994), the Best disease gene (VMD2) which is associated with Best macular dystrophy (Marquardt et al., 1998; Pertukhin et al., 1998), the RDS/peripherin gene responsible for a proportion of adult vitelliform macular dystrophy/butterfly dystrophy (Felbor et al., 1997; Kohl et al., 1998) and the EGF-containing fibrillin-like extracellular matrix protein-1 (EFEMP1) associated with Malattia Leventinese/Doyne honeycomb retinal dystrophy (Stone et al., 1999). However, TIMP3 (De La Paz et al., 1997), VMD2 (Krämer et al., 2000), EFEMP1 (Stone et al., 1999), and RDS/peripherin (Shastry and Trese, 1999) have been comprehensively analyzed and excluded as a major factors in the predisposition to AMD.

A genetic basis for AMD does not exclude an environmental influence on the disease process. Among the environmental factors, exposure to sunlight may be associated with the development of early AMD (Cruickshanks et al., 2001). In addition, several studies have reported a positive association between cigarette smoking and AMD (Hammond et al., 1996;). Several studies have found a positive association between risk factors for cardiovascular disease and AMD (Hyman et al., 2000; AREDS Group, 2000) while others have not (Klein et al., 2003). The Eye Disease Case Control Study (EDCCS Group, 1993) has demonstrated that higher serum carotenoid (beta carotene) levels seem to protect people from AMD. Among the specific carotenoids, lutein and zeaxanthin, which are primarily obtained from dark green, leafy vegetables, were most strongly associated with a reduced risk for AMD. The Macular Degeneration

Risk Factor Study has found less of the most serious type of AMD in patients with high antioxidant blood levels.

1.3.2 Oxidative stress and AMD pathogenesis

Evidence from a variety of studies suggests that RPE cells are susceptible to oxidative damage. From an anatomical point of view, RPE is located between the sensory retina and choroid, this location subsequently exposes RPE cells to a highly oxidative environment due to high oxygen partial pressure from the underlying choriocapillaris. Physiologically, RPE cells phagocytose and digest POS. Since the shed POS are extremely rich in PUFA, the oxidation of these PUFA initiates a chain reaction resulting in an abundance of reactive oxygen species (ROS), including lipid aldehyde radicals (Srivastava et al., 1995). During aging and pathological conditions, the balance between the ROS generation and the ROS clearance can be disturbed and result in oxidative damage to macromolecules (Ames et al., 1993; Harman, 1998). One of the most prevalent theories of aging is the mitochondrial theory. It proposes that oxidative damage to the mitochondria can lead to a spiral of confounding effects, whereby damaged mitochondria in turn release additional ROS, further increasing oxidative damage, and eventually leading to dysfunctional or defective mitochondria (Harman, 1981). Subsequently, cytochrome c is released into the cytoplasm, which initiates apoptosis through the activation of caspases that represents an early pathological event of AMD (Liang and Godley, 2003). The susceptibility of mitochondrial DNA to oxidative damage in the human RPE, together with the age-related decrease of cellular anti-oxidants, provides the rationale for a mitochondria-based model of AMD (for reivew, see Liang and Godley, 2003).

1.3.3 Growth factors and AMD

By virtue of its location, RPE forms the outer blood-retina barrier, which facilitates the transport of selective molecules between the outer neural retina and the choroidal blood supply. Among various growth factors secreted from RPE cells, it was demonstrated that differentiated RPE cells express high levels of vascular endothelial growth factor (VEGF) and pigment epithelium-derived factor (PEDF). A critical balance between VEGF and PEDF is important to prevent the development of choroidal neovascularization (CNV) in AMD (Ohno-Matsui et al., 2001) which is characterized by the growth of new blood vessels from the choroid through Bruch’s membrane into the subretinal pigment epithelial and subretinal space. These vessels can leak and bleed, leading to exudative retinal detachment and hemorrhage. Disturbed interactions of RPE cells with their surrounding extracellular matrix (ECM) could play a role in this process. The abnormal ECM may block normal signalling which could affect RPE adhesion and survival leading to degeneration and stimulating of angiogenic factors. A recent study reported that VEGF upregulates the production of PEDF by human RPE cells through VEGFR-1 in an autocrine manner (Ohno-Matsui et al., 2003). This regulation leads to the restoration of a normal balance between angiogenic stimulators and inhibitors, and this balance might paly a key role in maintaining the homeostasis of the human retina. Several therapeutic approaches have been taken to inhibit VEGF signalling (Presta et al., 1997; Krzystolik et al., 2002).

1.3.4 Role of apoptosis in AMD

Apoptosis is a genetically controlled mechanism of cell death characterized by a highly specific sequence of events that include cytoplasmic and nuclear condensation and fragmentation of nuclear chromatin (Wyllie et al., 1980). The mechanism of photoreceptor death is poorly understood, but one study found evidence of apoptosis of photoreceptors in 4 of 16 eyes with AMD using terminal deoxynucleotidyl transferase dUTP nick end-labeling (TUNEL) (Xu et al., 1996). A recent study reported that the number of TUNEL-positive cells in the choroid, RPE, outer nuclear layer (ONL), and inner nuclear layer (INL) is significantly greater in AMD eyes vs control eyes, suggesting that these cells may die by apoptosis (Dunaief et al., 2002). Moreover, the TUNEL-positive photoreceptors are clustered near areas of RPE atrophy and are mainly rods.

1.3.5 Vitamin A metabolism and AMD

Vitamin A and its derivatives play a central role in the vertebrate and invertebrate visual cycle (Saari, 2000). Retinoids function as the chromophores of the various visual pigments in animals, and photoisomerization of 11-cis-retinal to all-trans-retinal is the initiating event in vision. The role of the RPE in the visual cycle became more clear with the demonstration that the critical enzymatic regeneration of the 11-cis configuration occurred within this tissue (Bridges, 1976; Rando, 1996). Retinol uptake occurs at both the basolateral and apical surfaces of RPE by what appear to be separate receptor-mediated processes. The release of a crucial retinoid, 11-cis retinaldehyde (11-cis retinal), occurs across the apical membrane. Delivery of retinol across the basolateral membrane is mediated by a retinol binding protein (RBP) that is secreted by the liver as a complex with retinol (vitamin A) (Bok, 1993).

There is growing evidence for a key role for vitamin A in the formation of lipofuscin within the RPE. Rats maintained on a vitamin A deficient diet developed minimal autofluorecent material while rats raised on a high vitamin A diet accumulated significant amounts of lipofuscin-like fluorescence within the RPE. This has led to the conclusion that the accumulation of lipofuscin in rats is dependent on the dietary level of vitamin A (Katz et al., 1986).

2. Genetic approaches to correlate the genotype with phenotype

Two genetic approaches for revealing the genetic basis of human phenotypes are available, one that starts with a phenotype and concludes with the identification of a responsible gene or genes (*positional cloning approach*); the other that begins with a gene and works toward identifying one or more phenotypes resulting from allelic variation of it, this is called *candidate gene approach* (Dryja, 1997).

The first, the positional cloning approach (Collins, 1995), involves linkage analysis to identify the genetic region within which lies a disease-causing gene. Once the interval is of a manageable size, genes within the minimal region are tested as candidates for the disorder of interest. Over the last 15 years the technique of positional cloning has successfully been used to find disease genes. This technology has proven fruitful in the identification of genes for single gene disorders such as cystic fibrosis (Rommens et al., 1989), VMD2 (Marquardt et al., 1998;Petrukhin et al., 1998) and RS1 (Sauer et al., 1997). In positional cloning, the disease locus is mapped to a region of the genome using linkage studies in families affected with the disease. These regions are reduced in size using exclusion mapping techniques to give a small region of the genome (0.05 – 0.2%), in which to search for the gene. These single gene disorders are usually prevalent at low levels in the population, and are typically caused by a mutation in a single gene which results in an abnormal phenotype that is highly penetrant in family members who inherit the mutation. The allele frequency of these mutations in the population is less than 1%. Variants present in the population in greater than 1% are designated as polymorphisms, rather than mutations.

The second method, the candidate gene approach (Collins, 1995), involves examining individual genes that encode proteins with known functions or that have an expression profile that makes them candidates for the disorder in question. One factor that should be considered during the process of assessing any candidate gene is whether the gene is expressed in the tissue affected by the disorder. Alternatively, specifically or differentially expressed genes in a given tissue can be identified. Genes expressed in a certain tissue can be assumed to be important for the function of that tissue.

Both methods are typically used simultaneously to discover genes causing hereditary disorders. Reductions in the cost and improvements in the speed of scanning individuals for DNA sequence variations may make a gene-based approach an efficient alternative to phenotype-based approaches to correlating genes with phenotypes (Dryja, 1997). Both approaches require a group of affected subjects who have been reliably diagnosed with the disease. Candidate gene approaches also require a reliable control group without the disease so that a clear distinction between polymorphisms and true mutations related to disease can be made.

Many of the common diseases which affect millions of people, such as asthma and Alzheimer’s disease (AD), have an increased prevalence in certain families or populations. In these cases, the interaction of multiple disease susceptibility genes with environmental factors may lead to disease development. Asthma is a common respiratory disorder characterized by recurrent episodes of coughing, wheezing and breathlessness. Although environmental factors such as allergen exposure are risk factors in the development of asthma, both twin and family studies point to a strong genetic component. ADAM33 (a disintegrin and metalloproteinase domain 33) (Yoshinaka et al., 2002) represents one of the first putative asthma susceptibility gene identified through positional cloning in an outbred population (Van Eerdewegh et al., 2002). Confirmation for linkage has come from the separate analysis of the UK and US families, and from two separate, previously published, genome wide studies in other UK and US outbred populations (Van Eerdewegh et al., 2002). Physical mapping and direct cDNA selection identified 40 genes in the region under the peak of linkage. A survey of 135 polymorphisms in 23 genes identified the ADAM33 gene as being significantly associated with asthma using case-control, transmission disequilibrium and haplotype analyses.

Various appraoches have been developed to isolate and identify differentially expressed genes in tissues or cells. These include differential display (Liang and Pardee, 1992), RNA fingerprinting using arbitrarily primed PCR (McClelland et al., 1995), differential selection (Höög, 1991), and computer-based (in Silico) searches of public databases (Altschul et al., 1997), subtraction of cDNA libraries (Duguid and Dinauer, 1990), large scale generation of expressed sequence tags (ESTs) from cDNA libraries (Liew et al., 1994), serial analysis of gene expression (SAGE) (Velculescu et al., 1995), and suppression subtractive hybridization - SSH (Diatchenko et al., 1996 and 1999). However, none of these techniques are free of biases and are associated with a number of shortfalls. For example, SAGE although it is a very effective approach for determining the expression of mRNA populations, there are significant biases in the observed results that caused by sampling and sequencing error, nonrandomness and nonuniqueness of tag sequences (Stollberg et al., 2000). Suppression subtractive hybridization (SSH) is a technique that combines normaliation and subtraction. The abundant transcripts are suppressed, while rare transcripts are enriched to the same level of magnitude by normaliation. One of the objectives of using SSH technique is to isolate and identify genes that are enriched in the target tissue but not in the tissue or cell used as driver. This objective has the assumption that clones isolated by SSH are unique and only found in the target tissue, and therefore indicating their biological important. In addition, this might reveal a key physiological function specific of the tissue.

Over the last decade, different computational methods have been developed to identify genes and determine their functions. The following section reviews the existing approaches to predicting genes in eukaryotic genomes and underlines their advantages and limitations.

3. Computational methods for identifying genes

The identification of coding sequences (CDS) is an important step in the functional annotation of gene sequences. CDS prediction for mammalian genes from genomic sequence is complicated by the vast abundance of intergenic sequence in the genome, and provides little information about how different parts of potential CDS regions are expressed. Essentially, two different types of information are currently used to identify genes in a genomic sequences: (i) content sensors are measures that classify a DNA region into defined types, e.g. coding versus non-coding. (ii) signal sensors are measures that try to detect the presence of functional sites specific to a gene.

A sequence similarity-based approach is founded on sequence conservation due to the functional constraints and is used to search for regions of similarity between an uncharacterized sequence of interest and already characterized sequences in a public sequence database. A query sequence can be compared with DNA, protein, or EST sequences or it can be searched for known motifs. If a query sequence is found to be significantly similar to an already annotated sequence (DNA or protein) the information from the annotated sequence can be used to possibly infer gene structure or function of the query sequence. The basic tools for detecting sufficient similarity between sequences are local alignment methods ranging from the optimal Smith–Waterman algorithm to fast heuristic approaches such as FASTA (Pearson and Lipman, 1988) and BLAST (Altschul et al., 1990). Comparison with an EST database can provide information if the sequence of interest is transcribed, that is, contains an expressed gene, but will only give incomplete clues about the structure of the whole gene or its function, as ESTs only reflect a partial mRNA. Moreover, the correct attribution of EST sequences to an individual member in a gene family is not a trivial task.

The second computational approach, for the prediction of gene structures in genomic DNA sequences, termed the template approach, integrates coding statistics with signal detection into one framework. Coding statistics behave in a different manner on coding and non-coding regions and they are measures indicative of protein coding potential. This coding statistic is usually implemented as a 5th order Hidden Markov-Model (HMM). Signal sensors attempt to mimic closely processes occuring within the cell. They are intended to identify sequence signals, usually only several nucleotide-long subsequences, which are recognized by the cell machinery and are the initiation of certain processes. The signals that are usually modeled by gene-finding programs are promoter elements, start and stop codons, splice sites, and polyadenylation sites.

Singal sensors are not sufficient to elucidate gene structure, and it is necessary to combine them with coding statistics methods in order to obtain satisfactory predictive power.

A large body of literature on the subject of gene prediction has accumulated in the past 20 years. Early studies by a number of groups (Shepherd, 1981; Fickett, 1982; Staden and McLachlan, 1982) showed that statistical measures related to biases in amino acid and codon usage could be used to approximately identify protein coding regions in genomic sequences. Since then, numerous other compositional differences between coding and non-coding DNA sequences have been noted. Based on these differences, the *first generation* of gene prediction programs, designed to identify approximate locations of coding regions in genomic DNA were developed. The most known such programs are `TestCode` (Fickett, 1982), and `GRAIL` (Uberbacher and Mural, 1991), which uses a neural network approach to integrate multiple types of content statistics in order to classify sequence windows as coding or non-coding. These methods are generally able to identify coding regions of sufficient length, i.e. at least one or two hundred nucleotides, with fairly high reliability, but do not accurately predict precise exon locations. *Second generation methods,* such as `SORFIND` (Hutchinson and Hayden, 1992), `GRAIL II` (Xu et al., 1994) , use a combination of splice signal and coding region identification techniques to predict “spliceable open reading frames” i.e. potential exons), but do not attempt to assemble predicted exons into complete genes. *Third generation methods* attempt the more difficult task of predicting complete gene structures, i.e. set of exons which can be assembled into translatable mRNA sequences. *Fourth generation methods* of gene identification as exemplified by programs like `GENSCAN` (Burge and Karlin, 1997), has several significant advantages over existing gene finding algorithms. Predictive accuracy has been shown to be substantially higher for `GENSCAN` than for any other available method when tested on standardized sets of human and vertebrate genomic sequences. The program is able to identify 70 to 80% of exons in a genomic sequence precisely, with even higher levels of accuracy observed for complex genes containing ten or more exons (Burge and Karlin, 1997). Furthermore, a consistently high level of accuracy has been attained for sequencing of differing C+G% content. Other important features are the ability to treat partial as well as complete genes and the ability to predict multiple genes, occurring on either or both DNA strands, in a single sequence. These features make `GENSCAN` useful for analysis of the long genomic contigs which are being generated at an increasing rate by the Human Genome Project and related genome sequencing efforts. Another noteworthy feature of the program is its ability to assign a meaningful reliability measure, the exon probability to each predicted exon, which gives the user a highly informative guide as to the degree of confidence which should ascribed to each aspect of a prediction.

`TWINSCAN` is a new gene-structure prediction system that directly extends the probability model of `GENSCAN`, allowing it to exploit homology between two related genomes (Korf et al., 2001). Separate probability models are used for conservation in exons, introns, splice sites, and UTRs, reflecting the differences among their patterns of evolutionary conservation. `TWINSCAN` is specifically designed for the analysis of high-throughput genomic (HTG) sequences containing an unknown number of genes. In experiments on high-throughput mouse sequences, using homologous sequences from the human genome, `TWINSCAN` shows notable improvement over `GENSCAN` in exon sensitivity and specificity and dramatic improvement in exact gene sensitivity and specificity (Korf et al., 2001). This improvement can be attributed entirely to modeling the patterns of evolutionary conservation in genomic sequence (Korf et al., 2001).

The latest generation of gene prediction algorithms, such as `Grail/Exp`, `Geni EST`, `GenomeScan`, and `SGP2` combine *ab inito* prediction with similarity data into a single probability model. The basic idea of the `GenomeScan` program is to combine sequence similarity information, which can indicate the approximate location of many coding exons, with modeling of exon–intron and splice signal composition to aid in identification of additional exons and for determination of precise exon–intron boundaries (Yeh et al., 2001). `SGP2` is a new gene prediction program that combines ab initio gene prediction with TBLASTX searches between two genome sequences to provide both sensitive and specific gene predictions (Parra et al., 2003). The accuracy of `SGP2` when used to predict genes by comparing the human and mouse genomes is assessed on a number of data sets, including single-gene data sets, the highly curated human chromosome 22 predictions, and entire genome predictions from ENSEMBL (Parra et al., 2003). It was found that `SGP2` outperforms other ab initio gene prediction methods. `SGP2` was used to generate a complete set of gene predictions in both the human and mouse genomes. It was also reported that another few thousand human and mouse genes currently not in ENSEMBL (http://www.ensembl.org/) are worth verifying experimentally (Parra et al., 2003).

3.1 Public databases and sequence analysis tools

In recent years, development of the technology for efficient, automated DNA sequencing has led to the accumulation of large databases of DNA and protein sequences, and a new field of study known as "computational molecular biology" or "bioinformatics" has started to take shape as researchers work to interpret and draw conclusions from this wealth of new information. Though difficult to define, the field might be described as the area of research at the intersection of molecular biology, molecular evolution and structural biology which seeks to understand the relationships between sequence, structure, evolution and biological function by statistical/computational analysis of molecular sequences. In the past decade, with the shift in the emphasis of the Human Genome Project (HGP) from physical mapping to intensive sequencing, the problem of the identification of the precise exon-intron sturctures of genes in higher eukaryotic and especially human genomic DNA sequences has taken on significant practical importance.

The National Center for Biotechnology Information (NCBI) at the National Institutes of Health (USA) was created in 1988 to develop information systems for molecular biology. The main objective of NCBI was to maintain the GenBank (Benson et al., 2000) nucleic acid sequence database, to which data are submitted directly by the scientific community. In addititon, to provide data retrieval systems and computational resources for the analysis of Genbank data and various other biological data made available through NCBI web site. NCBI data retrieval tools include Entrez, LocusLink, PubMed and the Taxonomay Browser. Data analysis resources inlcude BLAST, Electronic PCR (e-PCR), OrfFinder, RefSeq, UniGene, HomoloGene, Database of Single Nucleotide Polymorphisms (dbSNP), Human Genome Sequencing, Gene Expression Omnibus (GEO), Online Mendelian Inheritance in Man (OMIM), Clusters of Orthologous Groups (COGs) database. The synchronization with the European Molecular Biology Laboratory (EMBL) Data

Library (Stoesser et al., 2003) and the DNA Data Bank of Japan (Tateno et al., 2002) provides comprehensive worldwide coverage.

3.1.1 Sequences analysis resources

3.1.1.1 BLAST

BLAST (Basic Local Alignment Search Tool) is a family of sequence alignment algorithms (Altschul et al., 1990), and contains a collection of programs with versions for query-to-database pairs such as nucleotide- nucleotide (BLASTN), protein-nucleotide (BLASTX), protein-protein (BLASTP), nucleotide-protein (BLAST). BLAST uses a heuristic algorithm that seeks local as opposed to global alignments and is therefore able to detect relationships among sequences that share only isolated regions of similarity (Altschul et al., 1990). Successful searches return a set of gapped alignments between the query and similar database sequences, with links to the full database records. Each alignment receives a score and a measure of statistical significane, called expectation value, for judging the alignment quality. BLAST is used to identify whether a given sequence is novel, homologous to a known sequence, or contains protein motifs which may provide clues regarding a role for the sequence being queried (http://www.ncbi.nlm.nih.gov/BLAST/).

An important and essential BLAST tool, BLAST2Sequences, compares two DNA or protein sequences and produces a dot-plot representation of the alignments in the form of a report (Tatusova and Madden, 1999). The NCBI generated assembly of both finished and unfinised human genome sequences can be searched through a specialized interface called Human Genome BLAST (http://www.ncbi.nlm.nih.gov/genome/seq/page.cgi?F=HsBlast.html&&ORG=Hs). Human Genome BLAST shows custom ‘Genome View’ of the BLAST hits which is integrated with the Human Genome MapViewer, showing confirmed and predicted gene location, or EST hits.

3.1.1.2 dbEST

dbEST is the EST division of NCBI and contains sequence and mapping data on "single-pass" short (about 300-600 bp) cDNA sequences or expressed sequence tags from a number of organisms. Usually produced large numbers, the ESTs represent a snapshot of the genes expressed in a given tissue, and/or at a given developmental stage. dbEST release 062703 (as of July 6, 2003) contains 17291123 entries. Human and mouse are the most widely represented organisms in dbEST, with 5 372 189 and

3 780 061 entries, respectively (as of July 6, 2003) (http://www.ncbi.nlm.nih.gov/dbEST/).

3.1.1.3 Online Mendelian Inheritance in Man (OMIM)

NCBI provides the online version of the OMIM catalog of human genes and genetic disorders (McKusick, 1998). The database contains information on disease phenotypes and genes, including extensive descriptions, gene names, inheritance patterns, map locations and gene polymorphisms (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=OMIM). OMIM is updated daily, it contains 14598 entries, including data on 10839 established gene loci and 1358 phenotypic descriptions (as of July 3, 2003).

3.1.1.4 dbSNP

In collaboration with the National Human Genome Research Institute (NHGRI), The NCBI has established the dbSNP database to serve as a central repository for both single base nucleotide subsitutions and short deletion and insertion polymorphisms (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Snp). There is no requirement or assumption about minimum allele frequencies or functional neutrality for the polymorphisms in the database. Thus, the scope of dbSNP includes disease-causing clinical mutations as well as neutral polymorphisms.

3.1.1.5 Pfam

Pfam is a **P**rotein **fam**ilies database of alignments and hidden Markov models (HMMs) covering many common protein domains and families. Pfam contains annotation of each family in the form of textual descriptions, links to other resources and literature references (http://pfam.wustl.edu/). Pfam version 10.0 (July 2003) contains alignments and models for 6190 protein families, based on the Swissprot 41.10 and SP-TrEMBL 23.15 protein sequence databases. Pfam stores an HMM profile constructed from a seed sequence alignment (Eddy, 1998) Pfam searches for matches to the HMMs. The Pfam database is invaluable for predicting the function of new sequences based on homology to previously characterised proteins.

3.1.2 Database Retrieval Tools

3.1.2.1 LocusLink

The LocusLink database of official gene names and other gene identifiers provides a single query interface to curated sequences and descriptive information about genes

(Pruitt and Maglott, 2001). LocusLink maintains descriptive information about loci including nomenclature, database identifiers, disease associations, map positions, sequence accessions, OMIM numbers, UniGene clusters, homology, map locations, and related web sites (http://www.ncbi.nlm.nih.gov/LocusLink/). The NCBI reference sequences (RefSeq) provide standards for complete genomic nucleic acids, assembled contigs, transcripts and protiens.

3.1.2.2 Source

Source is a web-based database that captures information from a broad range of resources, and provides it in manner particularly useful for genome-scale analyses (Diehn et al., 2003). Users can search for individual genes as well as simultaneously extract data for thousands of genes in batch. Source database can be searched using a gene symbol, the GenBank accession, the LocusLink identifier, or the UniGene cluster identifier. Source database also provides *in silico* generated expression calculated from EST abundance data. Other valuable features of the Source database include SwissProt functional information, GeneOntology annotations, Links to outside database such as LocusLink and UCSC Genome Browser, chromosomal location, the LocusLink descriptive summary, Aliases, and a link to Source’s microarray gene expression data (http://source.stanford.edu/cgi-bin/sourceSearch).

3.2 Genome analysis and annotation

The standard steps involved in the structure-function annotation of uncharacterized proteins include (1) sequence similarity searches using programs such as BLAST, FASTA; (2) identifying functional motifs and structural domains by comparing the protein sequence against PROSITE, BLOCKS, SMART, or Pfam; (3) predicted structural features of the protein, such as likely signal peptides, transmembrane segments, coiled-coil regions, and other regions of low sequence complexity; and (4) generating a secondary, and if possible tertiary, structure prediction.

Comparative analyses of several bacterial, archaeal, and eukaryotic genomes showed that the sequence comparison methods mentioned above failed to predict protein function for at least one-third of gene products in any given genome. In these cases, other approaches can be used that take into consideration all other available data, putting them into “genome context” (Huynen and Snel, 2000). With the availability of multiple complete genomes, the comparative approach is becoming the most powerful strategy for genome analysis.

It has been proposed that using association analysis of single nucleotide polymorphism (SNP) markers in candidate genes may be useful in identifying disease susceptibility genes for complex diseases. SNPs are plentiful throughout the human genome, being found in exons, introns, promoters, enhancers, and intergenic regions, allowing them to be used as markers. The next section provides insights into the SNPs and association studies.

4. SNPs as genetic markers

The focus in recent years on single nucleotide polymorphisms (SNPs), as initiated by groups such as the SNP Consortium and the U.S. National Human Genome Research Institute (NHGRI), has recognized that these single-base variations could enable novel approaches for elucidating complex, polygenic diseases, and potentially lead to the identification of new drug targets and diagnostic tests. Occurring on average once every 1000 bases along the human genome, SNPs might act as genetic markers linked to disease susceptibility genes, and coding SNPs in genes or control regions might even directly influence susceptibility to cancer, heart disease, diabetes, and other common diseases. In the postgenomics era, the focus has shifted to high-throughput studies intended to elucidate how genomic differences between individuals relate to their predisposition to disease, and how best to diagnose and treat those individuals with specific reference to their individual genetic makeup.

4.1 SNPs and association studies

Association studies have recently received a great deal of attention as a tool for detecting the genetic variation responsible for human common diseases. Unlike traditional linkage studies, which uses recombination information in large pedigrees, association methods use recombination information at the population level. Thus, association methods have greater power to detect small and moderate genetic effects than does linkage analysis (Risch and Merikangas, 1996). SNP markers are preferred over microsatellite markers for association studies, because of their high abundance along the human genome (SNPs with minor allele frequency >0.1 occur once every 600 bp) (Wang et al., 1998), their low mutation rate, and the accessibility of high-throughput genotyping. The power of association studies based on SNPs depends not only on the sample size and density of the marker map but also on many other factors, such as the age and frequency of the disease mutations and SNPs and the extent of linkage disequilibrium (LD) in the region.

4.2 Haplotype blocks and haplotype tag SNPs

Recent studies have shown that the human genome can be partitioned into blocks with limited haplotype diversity (Daly et al., 2001; Johnson et al., 2001; Patil et al., 2001; Dawson et al., 2002; Gabriel et al., 2002). In each block, a small fraction of single-nucleotide polymorphisms (SNPs), referred to as " haplotype tag SNPs (htSNPs)," can be used to distinguish a large fraction of the haplotypes. These tag SNPs can potentially be useful for association studies, in that it may not be necessary to genotype all SNPs in a given interval (Johnson et al., 2001).

Haplotype blocks, together with the corresponding htSNPs and common haplotypes determined by haplotype block partitioning algorithms, can be used in genome wide association studies, as well as in the fine-scale mapping of complex disease genes. First, a small number of samples (e.g., 10 or 20 individuals) are chosen to be genotyped for a dense SNP map in a region, and the haplotypes of these individuals are identified simultaneously. Second, an algorithm for haplotype block partitioning is employed, to identify haplotype block structure and the htSNPs based on the genotypes of the sampled individuals. Third, a larger number of samples (patients and controls) are genotyped only at these htSNP marker loci. Fourth, association studies are conducted using all the genotyped samples, with knowledge of the haplotype block structure. It seems that the above approach can significantly reduce the genotyping costs (Johnson et al., 2001).

4.3 Identification of new polymorphisms

A variety of techniques are available for the identification of new polymorphisms. One commonly used approach for SNP screening is to amplify genes of interest by PCR, scan the PCR products for the presence of DNA variants by confirmation-based mutation scanning methods, and then sequence positive PCR products. With a vast amount of human ESTs and genomic clones in the public domain, computer-based sequence alignment and clustering also provide a rich source for SNP identification.

Single-strand conformation polymorphism (SSCP) (Orita et al., 1989) analysis is one of the most widely used methods for mutation detection. In SSCP, DNA regions with potential polymorphisms are first amplified by PCR. Single-stranded DNAs are then generated by denaturation of the PCR products and separated on a nondenaturing polyacrylamide gel. A fragment with a single base modification generally forms a different conformer and migrates differently when compared with wild-type DNA.

Other alternative conformation-based mutation screening methods include denaturing high performance liquid chromatography (DHPLC) (O'Donovan et al., 1998). DHPLC detects polymorphisms by analyzing the DNA mobility of different heteroduplexes using chromatography in a slightly denaturing condition. The WAVE® DNA analysis system from Transgenomic, Inc. (San Jose, CA, USA) uses temperature-modulated hereroduplex analysis. The sample first is hybridized with wild-type DNA to form a mixture of homo- and heteroduplexes. The heteroduplexes can be separated from the homoduplexes by column chromatography at a temperature that partially denatures the mismatched DNA. The first gene subjected to dHPLC analysis was the calcium channel gene CACNL1A4 (Ophoff et al., 1996). Since then, more than hundred genes have been analysed.

II OUTLINE OF THE PRESENT RESEARCH

The goal of this project was to systematically identify and characterize genes expressed exclusively or abundantly in the RPE and to assess their contribution to AMD. The work described in this thesis is essentially interdisciplinary in nature in that, while the basic subject matter is biological and results of biological interest are obtained, techniques from other fields are used, including bioinformatic and computer tools. The aims of the present study were five-fold. Firstly, to analyse the domain of the project using standard systems analysis and design techniques to compile a suitable requirements specification for the design of a system. To assess the commercial database management systems (DBMSs) available and their suitability to the current project in order to select one. To construct a prototype of the system and test it. The system should have functions covering issue for clone, gene, expression, function, and to keep track of the status of searches similarity.

Secondly, to analyze and querry the individual of 2379 cDNA sequences derived from the suppression subtractive RPE (SSRPE) cNDA library against sequences deposited in GenBank and dbEST databases. In addition, storage and maintainace of the blast searches in the constructed relational database management system (RDBMS). A powerful approach to the analysis of expression in the RPE involves single-pass partial sequencing of clones from the SSRPE cDNA library together with the comparative analysis of the resuling ESTs with entries in public databases. This process leads to the establishment of a comprehensive catalogue of RPE-derived ESTs.

Thirdly, to carry out *in silico* expression and functional profiling of known and predicted RPE genes that identified during the course of this study.

Fourthly, to characterize the full-length cDNA sequences for 1-2 human orthologous RPE and/or retina genes.

Finally, generation of a SNP map for RPE01-D2 that will facilitate future association studies. SNP mapping conducts using a combination of DHPLC and direct sequencing.

III MATERIALS AND METHODS

1. cDNA library construction

The suppression subtractive hybridization (SSH) method was adapted to generate a cDNA library highly enriched for differentially expressed bovine RPE transcripts. A suppression subtractive RPE (SSRPE) cDNA library was performed in-house by Andrea Gerhig using a PCR-Select cDNA Subtraction kit (Clontech) according to the manufacturer's protocol (for review see Diatchenko et al. 1996 and 1999). In brief, tester cDNA was isolated from 50 ng bovine RPE poly(A)$^+$ RNA using SMART PCR cDNA synthesis kit (Clontech), and driver cDNA was prepared from an equal proporation of 150 ng bovine poly(A) RNA of heart and liver. A modified random 3' SMART CDS primer II (5′-AGCAGTGGTAACAACGCAGAGTACNNNNNNT GTGG-3′) was used for the first-strand synthesis reaction. The SMART II A Oligonucleotide (5′-AAGCAGTGGTATCAACGCAGAGTACGCGGG-3′), which has an oligo (G) sequence at its 3′ end, base pairs with the deoxycytidine stretch, creating an extended template. Reverse transcriptase (RT) switches templates and continues replicating to the end of the oligonucleotide. The cDNA was then amplified by long distance polymerase chain reaction (LD-PCR) using 5′ PCR primer II A (5′ AAGCAGTGGTATCAACGCAGAGT 3′).

Both tester and driver cDNAs were separately digested with RsaI to obtain shorter, blunt-ended molecules. Tester fragments were divided into two samples and ligated with two different adaptors; adaptor 1 (5′-CTAATACGACTCACTATAGGGCTCG AGCGGCCGCCCGGGCAGGT-3′) and adaptor 2R (5′-CTAATACGACTCACTAT A GGGCAGCGTGGTCGCGGCCGAG GT-3′). Each ligated tester sample was then hybridized with an excess of driver cDNA. The hybridized samples were mixed, and a second round of hybridization was performed with excess driver, followed by two rounds of PCR resulting in exponential amplification of the differentially expressed sequences. The PCR products were analyzed by agarose gel electrophoresis. Products from the secondary PCRs were ligated into pCRII vector using a TA cloning kit (Invitrogen), and transformed into TOP10F one shot competent cells, and selected by blue/white screening.

A total of 2379 independent cDNA clones (1002 in-house and 1377 by Lynkeus Biotech, Wuerzburg, Germany) were picked randomly, inoculated directly into 96-well microtitre plates containing 100 µl Luria-Bertani (LB) broth with ampicilin (50 µg/ml) and cultured overnight at 37 °C. These plates were then stored at – 80°C for further analysis.

2. EST amplification and sequencing

Plasmid DNA was used as a template for PCR amplification of the clones using the vector M13 forward primer (5′-CGCCAGGGTTTTCCCAGTCACGAC-3′) and M13 reverse primer (5′-AGCGGATAACAATTTCACACAGGA-3′). Briefly, PCRs were performed in a 25 µl reaction consisting of 2µl plasmid DNA (10-100 ng); 10 pmol of each M13 forward and M13 reverse primer; 1.25 mM dNTPs; 1 unit Taq DNA polymerase; 1 X PCR buffer; 1.0 or 1.5 mM $MgCl_2$. Thermal cycling was as follows: 94° C/5 min; 29 cycles of 94° C/30 sec; 65° C/30 sec; 72° C/1 min and a final extension at 72° C/5 min. After the PCR amplification, 5µl of the products were analysed in a 1% agarose gel. Because the RPE subtracted library was non-directionally cloned, the 2379 RPE ESTs were generated from either the 5′ or the 3′ end of the cDNA clones using nested primers either PCRII forward (5′-CTCGGATCCACTAGTAACGG-3′) or PCRII reverse (5′GCCGCCAGTGTGA TGGATAT-3′). The sequencing was conducted using the ABI Prism Ready Reaction Sequencing Kit and the ABI 310 automated sequencer (Perkin-Elmer, Norwalk, USA).

3. Bioinformatics

3.1 Design and Implementation of a Relational Database Management System (DBMS)

To store and manipulate the data a commercial relational database management system (RDBMS) and especially Microsoft Access was used. The optimized and efficient database engine of a commercial RDBMS makes the system modular and versatile. The SSRPE cDNA library information may be easily translated into normalized relational tables. Structured query language (SQL) statements and operations may be used to query these tables and derive composite information. This type of data storage scheme creates an extremely robust and efficient data repository that can be extended almost indefinitely.

3.1.1 Structured Systems Analysis and Design Method (SSADM)

The **S**tructured **S**ystems **A**nalysis and **D**esign **M**ethod (SSADM) is the most comprehensive of structured development methods currently available. It provides a framework and specific techniques to drive the process of computer systems specification and design (Meldrum et al., 1993). SSADM is designed to support the building of information systems, i.e. systems which have an underlying repository of data which must be accessed and updated. Logical Data Modelling (LDM) and Relational Data Analysis (RDA) are two of the central techniques of SSADM. LDM is used to define the data requirements for an application and consequently assumes a critical role in SSADM. RDA, also referred to as Third Normal Form Analysis, is used to validate the LDM. By applying the two techniques of LDM and RDA, SSADM ensures that the resultant data model is devoid of any unnecessary data duplications and thus supports system processing in an efficient and flexible manner. Both LDM and RDA are used in designing the RDBMS for the current project.

3.1.2 Logical Data Modelling (LDM)

The LDM is termed “logical” in that it models the underlying data requirements of an application regardless of how data will be processed and stored. LDM consists of a diagram (called the Logical Data Structure) with supporting textual definitions. The model defines the data requirements for an application by means of two key concepts: the entities (or groupings or data) inherent in the application; and the relationships (or associations) between the entities.

An entity represents a concept, object or thing which is relevant to the application being developed and about which information needs to be held by the application. Each of which needs to be individually identified by a unique reference or “key”. For example, a Clone processing application will probably have entities such as CLONE (identified by a key of Clone Name or ID) and GENE (identified by a key of Gene symbol or ID). An entity is represented on the Logical Data Structure using box with its name (in singular) appears inside the box.

Each entity consists of a grouping of attributes (or data items), which together provide a detailed definition of the object or concept represented by the entity. During design phase of the application development process, entities will be used to derive the groupings of data which will be physically stored e.g. tables, record tpyes, and files. For example, for the current RDBMS, it is likely that the “objects” about which data needs to be held will include: Clone, Gene, Clone – Gene, Exon, Category, Contig, Expression and Function.

Some attributes may appear in more than one entity, where an attribute is a foreign key or where the attribute is assuming a different role. However, it is widely recommended when considering and examining entities, to identify the attributes which comprise the entity. If no attributes can be identified, the validity of the entity may be in doubt.

An entity is deemed to contain a foreign key if it contains the primary key of another entity, provided this foreign key does not form part of the entity’s own primary key. A foreign key means that there is a relationship between the entity containing the foreign key and the entity for which the foreign key is the primary key. The entity containing the foreign key is the “Detail” entity in the relationship. Whereas, the entity for which the foreign key is the primary key is the “Master” entity. In the following example, entity Clone-Gene contains a foreign key to entity category:

Category	Clone - Gene
Category ID	Clone ID
Category name	Gene ID
	Contig ID
	___*Category ID___
	Library type
	Search similarity

Foreign keys are sometimes denoted by placing a marker usually an asterisk – against them.

A relationship represents an association between two entities which is of importance to the application being developed. It is critical that relationships are carefully and fully documented as they will be used, during design, to derive the access mechanisms which are implemented to retrieve the physically stored data. To identify the relationships, each pair of entities must be checked for the possibilty of the existence of a relationship. When a relationship has been identified, the degree of the relationship should be established. Most relationships will be one-to-many (1:M) or many-to-one (M:1). One-to-one (1:1) and many-to-many (N:M) relationships may appear at this stage of design, but they should be resolved later by the addition of link entities. If a one-to-one relationship is identified it should be examined closely to see if the two entities can be combined. There are few cases where a 1:1 relationship can exist if the two entities have different keys, different lives and are governed by different time periods. Many-to-many relationships are not permitted in SSADM, they contradict the Master/Detail concept.

3.1.3 Relational Data Analysis (RDA)

The Relational Data Analysis (RDA) with its emphasis on the detailed analysis of data will tend to produce data structures which are large, complex and devoid of any data duplication. Compared to the RDA data structure, LDM tends to produce data structures which are smaller (i.e. fewer entities), simpler (few relationships) and more tuned to the system’s processing requirements (Meldrum et al., 1993). However, because of this, the LDM may contain data duplication and may be less flexible in accommodating any future changes to system processing.

A relation is a grouping of logically-related attributes. The aim of RDA is to produce relations which are in Third Normal Form (TNF). Relations in TNF are devoid of any data duplication and represent the most optimum groupings of attributes for storage purposes. The underlying concept behind a relation is that data can be represented by means of a two dimensional table, comprising a number of rows and a number of columns (Table 2). Each column of the table is an attribute of relation (similar in concept to an attribute of an entity) and each row of the table represents an occurrence of the relation (similar in concept to the occurrence of an entity). An example of a relation holding clone details is shown in Table 2.

Table 2 Clone relation

Clone_Name	Clone_sequence	Clone_length	Clone_organism	Clone_Tissue
RPE03-D12	AGCGTGGTCGCGGCCGAGGTACCTCC	594	Bos taurus	RPE
RPE20-A12	GGAGAGCTCCCAACGCGTTGGATGCA	371	Bos taurus	RPE
RPE01-B01	TAGCGTGGTCNGCGGCCGAGGTACTT	548	Bos taurus	RPE
RPE20-B03	TAGCGTGGTCGCGGCCGAGGTACCG	363	Bos taurus	RPE
RPE03-B07	TCGAGCGGCCGCCCGGGCAGGTACC	571	Bos taurus	RPE
RPE08-C04	TTAGCGTGGTCGCGGCCGAGGTACTT	591	Bos taurus	RPE

The values for each row in a relation must be different to that of any other row. While it is likely that different rows will have the same values for particular columns, *no two rows can have the same values for all columns*. In the clone relation above, while the values of certain columns are the same for particular rows (i.e. clone organism and clone tissue), each row is different from all others (Table 2). Rather than show relations as table, the name of each relation is given, followed by a list of the attributes (i.e. column names) comprising the relation. Just as each occurrence of an entity is identified by means of the primary key, each row of a relation must also be uniquely identified by means of a primary key, which will consist of a single or set of column values. The primary key of a relation is denoted by underlining the relevant attribute(s) and placing it (them) at the top of the list of attributes.

Relation: Clone
Clone Name
Clone Sequence
Clone Length
Clone Organim
Clone Tissue

In RDA, normalisation is the process of producing a data structure containing the optimum grouping of attributes. An attribute is deemed to be "dependent" on the primary key if the value of the attribute can only be determined if the value of the key is known. To start the RDA process, the source material must be identified. The source material will be derived from the inputs and outputs of the required system. The next step is the conversion of the source data into Un-Normalised Form by listing all the attributes contained in the RDA source, as well as identifying the key of the Un-normalised relation. The RDA source must be analysed to identify all attributes contained within the source. When identifying the attributes on an RDA source, attributes which can be derived from other attributes on the source can usually be ignored.

The data model derived from the RDA is compared with the LDM. Any discrepancies between the two models must be considered with the objective of identifying any changes to the LDM which are needed to reflect the optimum grouping of attributes resulting from the RDA technique.

3.1.4 Physical Database Construction (MS Access)

This represents the last phase of database design and involves actually building the tables and relationships into the database (in this case MS Access). The process of physical database construction simply involves converting the logical database design into actucal tables, and determination of what data types will be used for specific attributes. Microsoft Access is a relational database management system (RDBMS). RDBMS is a program that facilitates the storage and retrieval of structured information on a computer’s hard drive.

In general, there are two basic approaches to developing information systems, first approach involves in-depth system analysis, design, and implementation. The second approach is done using rapid prototyping (in which analysis, design, and implementation are done iteratively). MS Access provides a number of features (such as graphical design tools, wizards, and high-level macro language) that facilitate rapid prototyping. Access application is a database that contains all objects necessary for users to work with data, as well as specific property settings and macros that automate data entry tasks. One of the well-known characteristic features of MS Access is the objects. Objects within the database, such as tables, queries, forms and reports (Table 3), are modified through their property settings to guide a user through completion of database tasks. Macros and Visual Basic procedures can automate the database so that forms and reports are displayed when needed. In addition, SQL can be used to construct queries that are then executed against a database, the result being a recordset. SQL is a standard language that can be used in a variety of databases, not just Access, with some minor modifications needed for each database program. The four most important SQL statements are *SELECT*, *DELETE*, *INSERT*, and *UPDATE* and they all are supported in Access.

Table 3 MS Access Objects and their functions

Object	Function
Table	used to store data.
Query	used to retrieve specific information from a table.
Form	used to enter data into a table in a database, also used to view and modify data.
Report	used for displaying and printing data in an easy-to-read format.
Page	used to show data on the Internet.
Macro	used to automate frequently performed database tasks.
Module	a program written in Visual Basic to automate and customize database operations.

3.2 BLAST searches

Before being assembled into clusters, sequences corresponding to the vector and the nested primers used to construct the SSRPE cDNA library were removed from each clone. This resulted in ESTs of approximately 300-600 bp of high quality sequence. Contig assembly program CAP3 (Huang and Madan, 1999) was used to assemble the high quality sequences into clusters and singletons. Nucleotide and protein BLAST programs at the National Center for Biotechnology Information - NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) were used for sequence homology searches in public databases. Annotations of possible protein-coding genes were performed and recorded in the constructed RDBMS for further investigations.

A first sequence search was conducted using BlastN against all entries in the non-redundant GenBank database. Sequences with E-values expected equaling zero were considered to identify known genes or to have partial similarity to known genes. Sequences with E-value close to zero were analysed for overall similarity that could be a different gene from the same family, chimeric clones, alternative splicing or presence of vector sequences. Sequences including repetitive fragments (e.g., LINE, ALU) were considered after eliminating the repetitive fragments and the Blast searches of non-repetitive part of sequences revealed an E-value equal to zero. The sequences that showed no match in the non-redundant database were subjected to another search with the BlastN against the dbEST and HTGs database. A second sequence search using normalized transcripts was performed using BlastN against the human genome draft sequence (http://www.ncbi.nlm.nih.gov/genome/seq/page.cgi?F=HsBlast.html&&ORG=Hs). The BLAST parameters used were as follows: Blastn program with an expect value of IE-50, default filter, and the default for the penality for a nucleotide mismatch is “-q –3”. But, in the present study a penality for a nucleotide mismatch of “-q –1” was used to increase the homology in our cross species comparisons.

Clones showing high homology with previously described sequences were considered to represent known genes. Clones with high homology with more than one gene were identified on the basis of highest homology of human origin. ESTs for which no homology was found were considered unknown.

Functional classification of the RPE known genes was undertaken by using the LocusLink, GeneOntology and Source database. Additionally, functional classification of the genes was based on the primary reported function obtained from the PubMed literature (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?CMD=search&DB=PubMed). Domain searches were performed using Pfam at Washington University in St. Louis public online search tools (http://pfam.wustl.edu/).

4. Identification and characterization of genes

4.1 RNA extraction and first strand cDNA synthesis

The RNA Clean system (Hybaid) was used to extract total RNA from frozen samples of human heart, brain, retina, RPE, lung, and placenta tissue. Residual genomic DNA was removed by the DNA free kit (Ambion). For RT-PCR experiments, 1–2 µg of total RNA was reverse transcribed using the SuperscriptTM preamplification system according to the supplier’s protocol (Gibco-BRL, Karlsruhe, Germany), and then used as a template for PCR. First cDNA strands were synthesized in 20µl reaction mixtures as follows:

1 µl Oligo-$(dT)_{12\text{-}18}$ [500µg/ml]
1-5 µg total RNA
up 12 µl DEPC-H2O
1µl 10 mM dNTP-Mix [each 10 mM dATP, dCTP dGTP, dTTP]

The reaction was incubated at 70°C for 15 minutes and then put on ice for 5 minutes. The following reagents were added:

4µl 5×1st strand buffer [250mM Tris-HCl, pH 8.3; 375 mM KCl, 15mM MgCl2]
2µl 0.1M DTT
1µl SuperScript™ II [200U]

Samples were incubated at 42°C for 52 minute, and were then inactivated by heating at 70°C for 15 minutes.

To normalize for variations in the amount of input RNA and synthesized cDNA from each tissue, routine control PCR reactions were conducted on 1 µl of newly synthesized first-strand cDNA with two different primer pairs amplifying the housekeeping gene ß-glucuronidase (GUSB). Due to its ubiquitious expression at low levels (Bracey and Paigen, 1987), GUSB was used for standardization of the cDNA targets. Oligonucleotides GUSB3 (5′-ACTATCGCCATCAACAACACACTGACC-3′) and GUSB4R (5′-GTGACGGTGATGTCATCGAT-3′) amplify a 198-bp fragment in cDNA and a 376-bp fragment in genomic DNA spanning exons 3 to 4. Amplification with GUSB6F (5′-GATCCACCTCTGATGTTCAC-3′) and GUSB7R (5′-CCTTTAGTGTTCCCTGCTAG-3′) results in a 454-bp fragment in cDNA spanning exon 11 and exon 12. Genomic DNA is not amplified due to a large intervening sequences of about 13 kb (Miller et al., 1990). 5 µl of the PCR reactions were electrophoretically separated on a 1% agarose gel stained with ethidium bromide and the band intensities were compared.

4.2 Polymerase Chain Reaction (PCR)

Specific primer pairs were designed using the DOS-based Oligo version 2.0 (NAR) program (Rychlik and Rhoads, 1989) and confirmed by employing the Primer3 program available at http://www-genome.wi.mit.edu/cgi-bin/primer/primer3 for the amplification of the desired DNA or cDNA fragment. The PCR reactions were carried out as described (Saiki et al., 1988) in a 25 µl reaction mixture containing 50 ng of template DNA and 10 pmol of each primer, 1.25 mM of dNTPs, 1 U of Taq DNA polymerase, 1 × PCR buffer; 1.0 or 1.5 mM MgCl2. Samples were denatured at 94° C for 5 min, and 30 cycles were performed as follows: denaturing at 94° C for 30 sec, annealing for 1 min at a temperature dependent upon the specific primer pair, elongation at 72° C for 1 min, followed by another elongation for 5 min. Then 5 µl samples were loaded on 1% agarose gels stained with ethidium bromide, and finally the DNA was made visible under UV light.

The annealing temperature T_m can be determined experimentally but is usually calculated applying the simple formula:

$$T_m = (4 X (G + C) + (2 X (A + T)\ ^\circ C$$

in which (G + C) is the number of G and C nucleotides in the primer sequence, and (A + T) is the number of A and T nucleotides. The annealing temperature for a PCR experiment is determined by calculating the T_m for each primer and using a temperature 1- 5 °C below this number.

4.2.1 PCR$_x$ Enhancer solution

The PCR$_x$ Enhancer system (Invitrogen) is an optimized buffer and cosolvent system that simplifies PCR amplification of problematic and/or GC rich templates. For example, the clone RPE06-C10 was found to show similarity to a single-exon gene that is not annotated in the Human Genome Sequence draft (Build 33). This intronless gene is located in a GC-rich region. GC rich regions can be extremely difficult to amplify and sequence. RPE06-C10 was amplified using 2X PCR$_x$ Enhancer solution using 3 pair of primers (RPE06-C10F1 5′-CTGACCCCTCTTGCCCCC-3′; RPE06-C10R1 5′-CGCCGTCCTCCACCACCT-3′; RPE06-C10F2 5′-GCCTCACGCA CAACCACATC-3′; RPE06-C10R2 5′- TAGCGGTAGTGGTAGCCCTCC -3′; RPE06 -C10F3 5′-GTGCTCTACCTAAACCGCCG-3′; RPE06-C10R3 5′-GCTCTG GGATGGGACAAAGG-3′) that were designed in an overlapping manner to cover the entire length of the single-exon gene.

4.3 Reverse Transcriptase (RT) PCR

RT-PCR is conducting using standard PCR (4.2) and first strand cDNA (4.1) as template and forward and reverse primers. Oligonucleotide RPE01-D2F(5′GCAGCA AAA GCAACAGCAGC-3′) and RPE01-D2R (5′-TCAGAGCAGGCAGGATTCGC-3′) and first strand retina cDNA were used to RT-PCR amplify the RPE01-D2.

4.4 Northern blotting

The Oligotex kit (Qiagen) was used to extract total RNA from frozen samples of bovine heart, liver, brain, retinal pigment epithelium, kidney and lung tissue. The RNA preparation was done under sterile and Rnase-free conditions. Residual genomic DNA was removed by the DNA-free kit (Ambion).

Norhtern blot analysis was performed with 7 μg of total RNA from bovine heart, liver, brain, retina, RPE, kidney and lung electrophoretically separated in the presence of formaldehyde at 55-75 Voltage for 3 hours. 10 X MOPS buffer (0.2 M MOPS, 50 mM NaOAc, 10 mM Na_2 EDTA; pH 7.0) is used as electrophoresis buffer. 7 μg RNA (relevant tissue quantity) and 12 μl loading buffer (1 X MOPS, 18.5 formaldehyde, 50% formamide, 0.04% bromphenol blue, 10 μg ethidiume bromide) are mixed and denatured for 10 min at 65°C. Total RNA is vacuum transferred to Hybond-N^+ membranes (Amersham Pharmacia Biotech, Freiburg, Germany) in 20 X SSC buffer (3 M NaCl, 0.3 M sodium citrate; pH 7.0).

Selective individual clones of the SSRPE cDNA library were used as probes. The probes were purified and labeled with α^{32} P-dCTP by nick translation (Nick Translation System, Gibco BRL) according to the manufacturer’s protocol. The filters were hybridized overnight in 0.5 mM sodium phosphate buffer, pH 7.2; 7% SDS, 1 mM EDTA at 60^0 C. The hybridization buffer was removed and the filter was washed with washing buffer at 60^0 C for 15 min in 2X SSPE/0.1% SDS, 1X SSPE/0.1% SDS and 0.5 X SSPE/0.1% SDS (20 X SSPE: 3 M NaCl, 200 mM NaH_2PO_4, 20 mM Na_2 EDTA). The filters were wrapped with prafan membrane and exposed to a X-rays films at -80^0 C for 3 – 5 days.

4.5 Cloning of PCR products for RPE01-D2

The PCR products were ligated into pGEM® T-easy vector (Promega) and then transformed into **E. Coli** according to the manufacturer's protocol.

The LB medium and SOC were prepared as follows:

LB medium (per liter)

10g	Bacto®-Tryptone
5g	Bacto®-Yeast Extract
5g	NaCl

The pH was adjusted to 7.0 with NaOH

SOC medium (100 ml)

2g	Bacto®-Tryptone
0.5g	Bacto®-Yeast Extract
1ml	NaCl
0.25ml	1M KCl
1ml	2M Mg^{2+} stock, filter sterilized
1ml	2M glucose, filter sterilized

15g of agar was added to 1 liter of LB medium and then autoclaved. The medium was allowed to cool to 50°C before adding ampicillin to a final concentration of 100µg/ml. 30-35ml of the medium was poured into 85mm petri dishes. The agar was left to be hardened. The medium was stored at room temperature.

4.5.1 Ligation of RPE01-D2

The ligation reaction in a final volume of 10 µl was set up in 0.5ml tube (known to have low DNA-binding capacity) as follows:

5 µl	2 X Rapid ligation Buffer (it is important to vortex)
1 µl	pGEM-T Easy Vector (50ng)
1 µl	T_4 DNA ligase
3 µl	PCR product

The reaction was mixed by pipetting and then incubated overnight at 4°C.

4.5.2 Transformation of the ligated RPE01-D2

The tube containing the ligation reaction was centrifuged to collect contents at the bottom of the tube. 1µl of ligated DNA was added to the sterile 1.5ml microcentrifuge tube containing the thawed competent cells (JM109) on ice. The contents of the 1.5ml tube were put into cuvette tube. The cuvette was then put in the gene pulser (BIO-RAD), which was already turned on and adjusted to 2.50. The cells were heat shocked inside the cuvette by pressing the two buttons of the gene pluser until a sound was heard. 500µl of SOC medium was added to the cuvette tube. The contents of the cuvette tube were then transferred into 1.5ml tube. The 1.5ml tube was incubated for 1 hour at 37°C. Prior to plating, 20µl of 50mg/ml X-Gal and 100µl of 100mM IPTG were added to 2 LB/ampicilin plates. 100µl of transformation culture was plated onto the duplicate LB/ampicilin/IPTG/X-Gal plates. The plates were incubated overnight at 37°C. The unused portion of the transformation culture was stored at 4°C. Using blue/white screening, positive clones were recognized and selected. 30 0.5ml tubes were prepared with the mix of 20µl (4µl ampicilin+1ml LB medium). Each of the 0.5ml tube was innoculated with a colony. The 30 tubes (clone1 to clone30) were then incubated either at 37°C for 3 hours or overnight at room temperature.

4.5.3 Clones amplification and sequencing

PCR amplification of clones from Plasmid DNA was done using either the gene specific primers (GSP) RPE01-D2F and RPE01-D2R or the vector M13 forward primer and M13 reverse primer. 5 µl of the PCR products were separated by gel electrophoresis and visualized under UV illumination. PCR products showing inserts were treated with shrimp alkaline phosphatase (USB, Cleveland, USA) and exonuclease I (USB, Cleveland, USA) and then sequenced using the ABI Prism Ready Reaction Sequencing Kit and the ABI 310 automated sequencer (Perkin-Elmer, Norwalk, USA).

5. SNP genotyping

The genomic sequence as well as 5 kb upstream and downstream of the RPE01-D2 (official nomenclature RDH12) gene were downloaded from the NCBI Blast. The repeat mask web server program (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker), was used to determine the repeat regions of the genomic sequence of the RDH12 gene. The primers were designed to flank 25 fragments of the gene, using the DOS-based Oligo version 2.0 (NAR) program (Rychlik and Rhoads, 1989) and a web-based primer3 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3www.cgi/).

5.1 SNPs identification

A combination of denaturing high performance liquid chromatography (DHPLC) (O'Donovan et al., 1998) and direct DNA sequencing were used to screen the *RDH12* gene for single nucleotide polymphism (SNP). The genomic DNA was isolated from peripheral blood leukocytes of 16 control DNAs according to standard protocols. The DNA controls from 16 individuals were used to PCR amplify the 25 fragments of the *RDH12* gene. The individual coding exons, flanking intronic sequences as well as the 5′ and 3′ untranslated regions (UTR) of the *RDH12* gene were PCR amplified with the oligonucleotide primers and conditions listed in Table 4. Reactions were conducted for 33 cycles in 1.0 or 1.5 mM $MgCl_2$ containing buffer with or without 4% formamide and a touch-down protocol between 58° and 52°C (Table 4).

DHPLC analysis was performed on a Wave DNA Fragment Analysis System (Transgenomic Inc., Omaha, USA) In brief, 5-10 µl of the PCR products were denatured at 95°C for 5 min and cooled down to 65°C with a temperature ramp of 1°C/min to produce heteroduplex molecules in case of heterozygous DNA samples. DNA was eluted at a flow rate of 0.9 ml/min within a linear acetonitrile gradient consisting of buffer A (0.1 M triethylammonium acetate, TEAA) and buffer B (0.1 M TEAA, and 25% acetonitrile). Temperature selection for the successful heteroduplex separation in heterozygous fragments was carried out using the WaveMaker software (Transgenomic) typically within the range of 54 to 68°C. To identify homozygous mutations, aliquots of a known wild-type sample were added to the DNA prior to the re-annealing step to enable heteroduplex formation. The average analysis time per sample took 8-12 min including a regeneration and equilibration step. The linear acetonitrile gradient was adjusted to a retention time of the DNA peak at 4-5 min. PCR products with heteroduplex formation were subjected to sequence analysis in order to reveal the nature of the polymorphism.

5.2 Determination of allele frequency of the identified SNPs

The fragments that showed significant SNP frequency equal to or more than 20%, were then PCR amplified from 48 family DNA controls, purified and then were sequenced at MWG (MWG Biotech, München, Germany). The SeqMan™ II software (DNASTAR Inc., 1989-2002) (http://www.dnastar.com/cgi-bin/php.cgi?_r13.php) was used to align the sequences of the 48 DNA controls to determine the allele frequency for each identified SNPs. As a result of this, a SNP card was prepared for the *RDH12* gene, showing all the SNPs identified during the process of SNP mapping. Using the SNP card, the SNPs that are running together were determined from the 48 family DNA controls. These SNPs are in strong linkage disequilibium.

Table 4 Oligonucleotide primers and conditions of SNP genotyping for RDH12

Fragment number	Primer name	Primer sequence (5′- 3′)	ANNEALING TEMPERATURE °C	$MgCl_2$ mM*
1	ID2-5'UTR-F1	TGTCAATAGTGCCTGCTGTG		
	ID2-5'UTR-R1	CTTGCTGATTTGACCTTTGG	52	1.0-
2	ID2-5'UTR-F2	AGGGATAACCAGGAGACCAG		
	ID2-5'UTR-R2	CATCACCCACTTCATCATTGT	55	1.0-
3	ID2-5'UTR-F3	TTATGGCTTTGCTTATGTGC		
	ID2-5'UTR-R3	GCAAAAAGACCCTGATAACC	58	1.0-
4	ID2-5'UTR-F4	AAATTTAATGTTTATTTACTTAA		
	ID2-5'UTR-R4	GGAAACTGAAAACTAAAACT	52	1.0-
5	ID2-5'UTR-F5	GCCTTGTGTATGATGGTTTT		
	ID2-5'UTR-R5	ACCAACAGAACAAGGAGTGAT	58	1.0-
6	ID2-5'UTR-F6	GAAGTGTCTGCTGGGAATGA		
	ID2-5'UTR-R6	TGGTGTAGAAAAGGGAGAGA	58	1.0-
7	ID2-5'UTR-F7	GCAAGTGAGATAGCAAGGGA		
	ID2-5'UTR-R7	TAAAATAGGATGGGAAGGAA	56	1.0-
8	1D2-Exon1F	AGTAGAGGTGGCAGTGGTTG		
	1D2-Exon1R	GATGCTTCCTTCTGGTTTTC	58	1.0-
9	1D2-Exon2F	TGATTATTTGTGGCTTCTGG		
	1D2-Exon2R	GGTTCCCAGGTTTTACATTC	56	1.0-
10	1D2-Exon3F	CTACTGTGAAAAGCCCGAAG		
	1D2-Exon3R	CCAGCAGCACAACTTCATCT	58	1.0-
11	1D2-Exon4F	GAGATAGGTCCAAATGAAGG		
	1D2-Exon4R	ATGTAGATGTGACCCCTCCA	52	1.0-
12	1D2-Exon5F	AATCCACAAACTCAGACCAA		
	1D2-Exon5R	CAAATGAGATAAGAGATAAGATGT	58	1.0-
13	1D2-Exon6F	TTTGGAACATAGAAGGCTGAG		
	1D2-Exon6R	GCATAACCAACAGCGACAGT	52	1.0-
14	1D2-Exon7F	AAATCTGGAGGGCTTGGTCT		
	1D2-Exon7R	TCAGAGCAGGCAGGATTCGC	58	1.0-
15	1D2-3'UTR-F1	TTGTGAGACTGGCTTATGGC		
	1D2-3'UTR-R1	TGCTTTTTCTCTGTCTGCCT	58	1.0-
16	1D2-3'UTR-F2	AGAACTCAGGGCAAAGACAG		
	1D2-3'UTR-R2	AGCACCTGAACACCACGA	55	1.0-
17	1D2-3'UTR-F3	GTCGTTCCCCTTGTTCAGAT		
	1D2-3'UTR-R3	TTTCCTTTAGTTTCCCATTG	55	1.0-
18	1D2-3'UTR-F4	AATCTTTTTCTTTTGGCTCA		
	1D2-3'UTR-R4	TTTCAATACCCAATACCCAA	56	1.0-
19	1D2-3'UTR-F5	CTCCTCTCCCTCTACCATTG		
	1D2-3'UTR-R5	GTTTGAGGTATGCTTTTTGGA	58	1.0-
20	1D2-3'UTR-F6	CATCCAAAAAGCATACCTCAAAC		
	1D2-3'UTR-R6	AATGAAGCATACAGCTAGGC	58	1.0-
21	1D2-3'UTR-F7	GCCTAGCTGTATGCTTCATT		
	1D2-3'UTR-R7	GATAATAAGTTGAGAGTGGTGAC	58	1.0-
22	1D2-3'UTR-F8	GTCACCACTCT CAACTTATTATC		
	1D2-3'UTR-R8	TGATTTATCCAAGTATATACGTG	-	-
23	1D2-IVS1F1	TGCCTCAGCCTCCCAAGTAG		
	1D2-IVS1R1	AAATGCTGGAGTCAGGGCCA	-	-
24	1D2-IVS4F1	CCACGGAGGTAGGCAATC		
	1D2-IVS4R1	CAAAGCAAGAGCCAGAGC	56	1.0-
25	1D2-IVS6F1	CTGTCGCTGTTGGTTATGCC		
	1D2-IVS6R1	TTATCCATTCCCCTCTCATC	56	1.0-

*PCR reaction without (-) 4% formamide

IV RESULTS

1. RDBMS design and construction

This section provides insight into design and implementation of RDBMS. SSADM method have been used in designing and analysis of the RDBMS for the current project. MS Access was used in implementing the RDBMS. Utilizing the LDM, the entities for the RDBMS had been identified and analysed. Having identified a possible entity, the validity of the entity should be confirmed by determining the attributes which constitute the entity. If it is not possible to allocate any such attributes, the validity of the entity may be in question. The entities and their possible attributes for the current application are shown in Table 5.

Table 5 The identified entities and their possible attributes

Entity	Possible attributes
• **CLONE**	**Clone name**, Clone length, Clone sequence, PCR band size, Clone source, Clone organism, Clone tissue
• **GENE**	**Gene ID,** Gene name, Accession number, CDS, Gene length, Cytogenetic map, LocusLink ID, UGCluster, Number of exons, Aliases, Protein accession number
• **CLONE – GENE**	**Clone name, GeneID**, *Contig ID, Category ID, Subcategory ID,* Library type, Search similarity, % of similarity, Genomic Contig, Similarity with respect to CDS, DNA source
• **EXON**	**ExonID, GeneID**, Start of exon End of exon, Size of exon, Size of intron, Nucleotide sequence of the exon, Amino acid sequence of exon, Exon F-primer, Exon R-primer
• **CONTIG**	**ContigID,** Contig length Consensus sequence of Contig
• **EXPRESSION**	**GeneID,** Expression Assays, Probe, Transcript size (kb), Tissues panel, Tissue specificity
• **FUNCTION**	**GeneID,** Accession Number, Class function , Subclass function, Protein motif, Protein domain, Biology of the gene

A grid or matrix is constructed to aid the identification of relationships between the identified entities (Table 6). The grid is constructed by listing the entities along each of the two axis. The purpose of the grid is to ensure that each entity is considered in relation to every other entity so that relationships are not overlooked. To identify the relationships, each pair of entities must be checked for the possibilty of the existence of a relationship. When a relationship has been allocated, the degree of the relationship should be established. Most relationships will be one-to-many (1:M) or many-to-one (M:1). One-to-one (1:1) and many-to-many (N:M) relationships may appear at this stage of design, but they should be resolved later by the addition of link entities. If a one-to-one relationship is identified it should be examined closely to see if the two entities can be combined. Few cases where a 1:1 relationship can exist if the two entities have different keys, different lives and are governed by different time periods. Many-to-many relationships are not permitted in SSADM, they contradict the Master/Detail concept. In the current RDBMS the relationships between the entities were identified, and consequently an entity/entity matrix (or grid) was drawn up as shown in table 6.

Table 6 Entity/Entity Matrix for the current RDBMS*

Entity/Entity	**Gene**	**Exon**	**Clone**	**Expression**	**Function**	**Contig**
Gene		M:1	M:1	1:1	M:1	
Exon	1:M					
Clone	1:M					1:M
Expression	1:1					
Function	1:M					
Contig			M:1			

*1:M means one-to-many relationship, whereas M:1 refers to many-to-one relationship

A relationship must have a unique identifier with the following structure:

<Subject entity name> <relationship link phrase> <object entity name>

The relationship link phrase describes the entity from the perspective of the subject entity. A relationship line must be able to read from either end without any ambiguity. It is formulated as a plain English “Relationship Statement”, quoting optionality, meaning and degree. For example, the relationship statements for Figure 3, read:

A. Each GENE must be made up of one or more Clones. And each CLONE must be part of one and only one Gene.

B. Each GENE must be made up of one or more Exons. And each EXON must be part of one and only one Gene.

C. Each CONTIG must be made up of one or more Clones, and each CLONE must be part of one and only one Contig.

When referring to relationships, it is convenient to have a way of referring to the entity at the ´one´ end and the entity at the ´many´ end. The entity at one end is referred to as the `Master` entity in the relationship, and the entity at the many end is the `Detail` entity. For example in the Figure 3A, the entity Gene is the master entity in the relationship, the entity Clone is the detail entity.

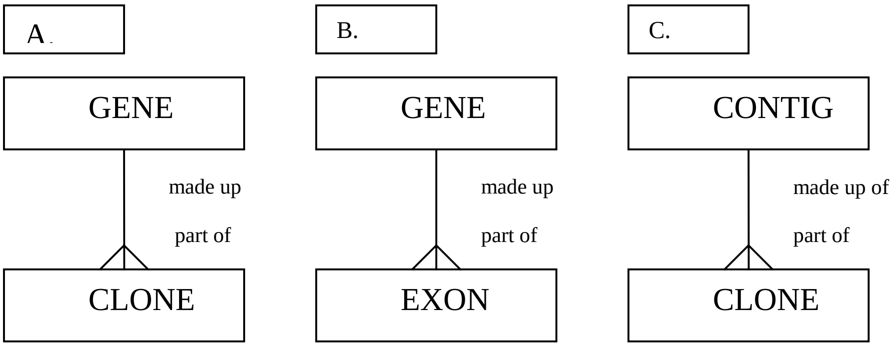


Figure 3 One-to-many relationship. For example, the relationship in A can be read as follows: each gene must be made up of one or more clones and each clone must be part of one and only one gene. The same degree of relationship can be applied to B and C.

Having the entities, attributes and relationships being allocated, a logical data model is created which models the information requirements of the current system, regardless of how this information is going to be processed (Figure 4).

A major emphasis of RDBMS is to provide reliable identification of RPE ESTs and to group them into clusters that represent transcripts from the same gene (Figure 5A). This clone duplication leads to data redundancy that needs to be normalized. For example, sequencing of 1002 ESTs yielded 46 clones that have identity with B. taurus retinal pigment (RPE1) (Figure 5B). The storage of RPE1 many times as well as its accession number and coding sequence (CDS) leads to data redundancy. Therefore, the data was split into two relations i.e. clone relation and gene relation. The aceession number is then used as a primary key to relate the two relations (Figure 5C).

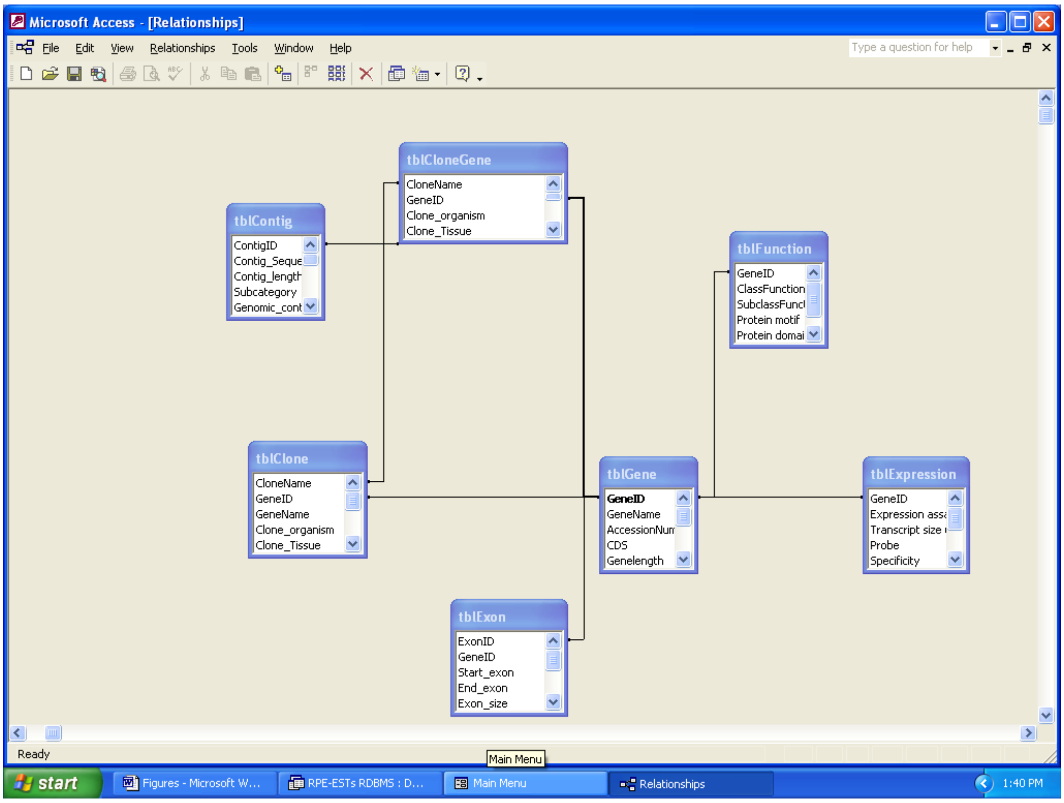
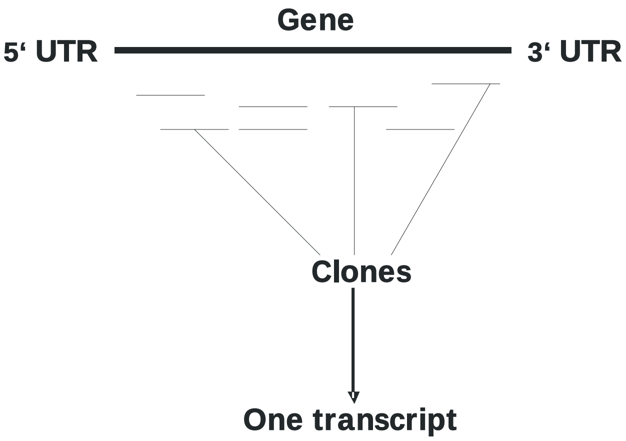


Figure 4 Logical Data Model showing the entities and their relationships created for the RPE-EST RDBMS. Each box represents an entity as summarized in Table 5.

Using the RDA, the complete normalization table for the RDMBS was created as shown in Table 7. The process of converting Un-Normalised Form (UNF) relations into First Normal Form (FNF) relations involves the identification and removal of repeating groups of attributes from the UNF relation. A repeating group comprises one or more attributes for which multiple values may be present in a single instance of the UNF primary key. All repeating attributes must be removed from the UNF relation to a separate relation that has as its key, the key of the UNF relation, as well as additional attribute (or attributes) which will uniquely identify an occurrence of the repeating group. FNF relations are therefore relations which do not contain any repeating groups. For example, for some genes in our dataset there is a repeating group of data about clones. In addition, for each contig there is a repeating group of data about clones. In these cases, the original un-normalized relation is split into three relations in the FNF (Table7).

A. Clone redundancy and normalization



B. Data redundancy

Plate ID	Clone ID	Gene Name	Accession Number	Coding Sequence CDS
RPE01	A03F	Bovine retinal pigment (RPE1) mRNA, 3' end Length = 1584	M81193	1..1477
RPE02	C11F	Bovine retinal pigment (RPE1) mRNA, 3' end Length = 1584	M81193	1..1477
RPE02	F06F	Bovine retinal pigment (RPE1) mRNA, 3' end Length = 1584	M81193	1..1477
RPE16	E12F	Bovine retinal pigment (RPE1) mRNA, 3' end Length = 1584	M81193	1..1477
RPE22	B08F	Bovine retinal pigment (RPE1) mRNA, 3' end Length = 1584	M81193	1..1477

C. Data normalization

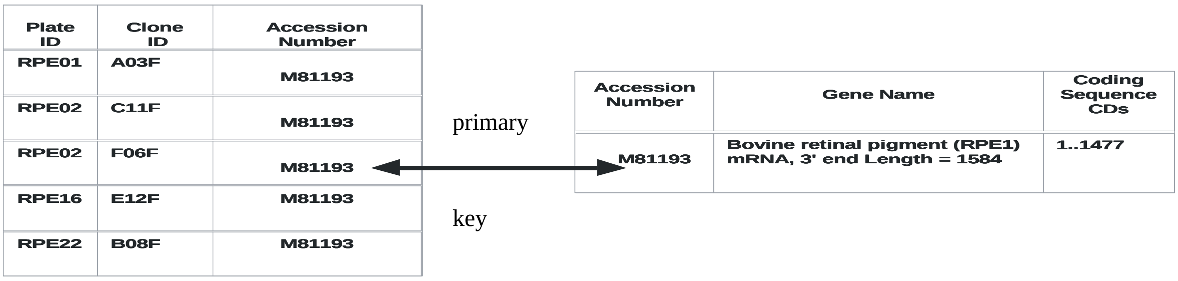


Plate ID	Clone ID	Accession Number
RPE01	A03F	M81193
RPE02	C11F	M81193
RPE02	F06F	M81193
RPE16	E12F	M81193
RPE22	B08F	M81193

Accession Number	Gene Name	Coding Sequence CDs
M81193	Bovine retinal pigment (RPE1) mRNA, 3' end Length = 1584	1..1477

Figure 5 Clones/data redundancy and normalization. (A) In our RPE cDNA library some genes are represented by more than one clone (1:M), this clone redundancy was normalized by considering those clones from the same gene (identical or nonoverlapping) as corresponding to one transcript; (B) Gene name, like RPE1, is repeated many times as well as accession number and coding sequence (CDS); (C) Data split into two relations; clone relation and gene relation. The clone relation comprises of Plate ID, clone ID and accession number, whereas the gene relation contains the gene name, accession number, and CDS.

To transform FNF relations into Second Normal Form (SNF) relations, each FNF relation with a compound key must be examined. Each attribute which forms part of the key of an FNF relation must be considered to determine whether the key would still uniquely identify all occurrences of the relation if the attribute in question were to be excluded. If such an attribute is identified, it should be removed from the primary key of the relation, but retained as a non-key attribute. For example, data about exons are determined by Exon ID and Gene ID alone i.e. given the value of Gene ID and Exon ID the other values about exons can be determined (Table 7).

The step of converting SNF relations into Third Normal Form (TNF) relations involves the identification and documentation of inter-data dependencies. Each pair of non-key attributes in SNF relation needs to be examined to identify whether there is any dependency between the two attributes. If inter-data dependency is identified, the dependent attribute is removed to a separate TNF relation as a non-key attribute. The primary key of this new TNF relation is the attribute on which the removed attribute is dependent. The primary key of the new TNF relation becomes a foreign key in the SNF relation from which the TNF relations has been created. For example, the expression and functional data about the gene has nothing to do with the other gene informations like CDS and number of exons. So, both expression and function are separated into a new relation (Table 7), but are related to the gene relation by a foreign key which in this case is the GeneID.

In order to validate the LDM with the results of RDA, the TNF relations must be converted into a data structure which can be compared with the LDM. An entity box is drawn for each TNF relation. Each entity box must be named and in each box, the relation’s primary and any foreign keys are shown.

Figure 6 shows the tables that were designed and implemented for the current RPE-ESTs RDBMS. These included tables corresponding for each entity identified for the system as shown in Table 5. Seven tables for entities GENE, CLONE, CONTIG, EXON, EXPRESSION, FUNCTION, and GENE-CLONE were designed and developed.

Table 7 Complete normalization table for the current RDBMS

UNF (Un-normalised Form)	**FNF** (First Normal Form)
Clone name	**Gene ID**
Gene name	Gene name
LocusLink ID	LocusLink ID
Protein accession number	Protein accession number
Aliases	Aliases
Clone organism	Accession number
Clone tissue	Coding sequence CDS
Clone source	Number of Exons
Library type	Exon ID
PCR band Size (bp)	Start of Exon
Clone Sequence	End of Exon
Clong length	Length of Exon
Gene ID	Exon Size (bp)
Accession number	Nucleotide sequence of Exon
Coding sequence CDS	Amino Acid sequence of Exon
Number of Exons	Exon F-primer
Exon ID	Exon R-primer
Start of Exon	Expression assays
End of Exon	Transcript size (kb)
Length of Exon	Probe
Exon Size (bp)	Tissues panel
Nucleotide sequence of Exon	Tissue specificity
Amino Acid sequence of Exon	Class function
Exon F-primer	Subclass function
Exon R-primer	Protein motif
% of homology	Protein domain
Place of homology – Query	Biology of the gene
Place of homology – Subject	
Similarity DNA Source	**Contig #**
Category ID	*_Category ID_
Similarity with respect to CDS	Contig consensus sequence
Comments/Remarks	Contig length
Expression assays	
Transcript size (kb)	**Clone name**
Probe	**Gene ID**
Tissues panel	*_Contig #_
Tissue specificity	*_Category ID_
Class function	% of homology
Subclass function	Clone organism
Protein motif	Clone tissue
Protein domain	Clone source
Biology of the gene	Library type
Genomic Contig	Clone sequence
Human Genomic Similarity	Clone length
Contig_Percentage of homology	PCR band Size (bp)
Contig_Place of Homology – Query	Place of homology – Query
Contig_Place of Homology – Subject	Place of homology – Subject
Contig_commnets/Remarks	Similarity DNA Source
Subcategory	Similarity with respect to CDS
Chromosome ID	Genomic Contig
Cytogenetic map	Human Genomic Similarity
Contig #	Contig_Percentage of homology
Contig consensus sequence	Contig_Place of Homology – Query
Contig length	Contig_Place of Homology – Subject
	Contig_commnets/Remarks
	Subcategory
	Chromosome ID
	Cytogenetic map
	Comments/Remarks

Table 7 (continued)

SNF (Second Normal Form)

Gene ID
Accession number
Gene name
LocusLink ID
Protein accession number
Aliases
Coding sequence CDS
Number of Exons
Expression assays
Transcript size (kb)
Probe
Tissues panel
Tissue specificity
Class function
Subclass function
Protein motif
Protein domain
Biology of the gene

Exon ID
Gene ID
Start of Exon
End of Exon
Length of Exon
Exon Size (bp)
Nucleotide sequence of Exon
Amino Acid sequence of Exon
Exon F-primer
Exon R-primer

Clone name
Clone organism
Clone tissue
Clone source
Library type
Clone sequence
Clone length
PCR band Size (bp)

Contig #
**Category ID*
Contig consensus sequence
Contig length

Clone name
Gene ID
**Contig #*
**Category ID*
% of homology
Place of homology – Query
Place of homology – Subject
Similarity DNA Source
Similarity with respect to CDS
Genomic Contig
Human Genomic Similarity
Contig_Percentage of homology
Contig_Place of Homology – Query
Contig_Place of Homology – Subject
Contig_commnets/Remarks
Subcategory
Chromosome ID
Cytogenetic map
Comments/Remarks

TNF (Third Normal Form)

Gene ID
Accession number
Gene name
LocusLink ID
Protein accession number
Aliases
Coding sequence CDS
Number of Exons

Gene ID
Expression assays
Transcript size (kb)
Probe
Tissues panel
Tissue specificity

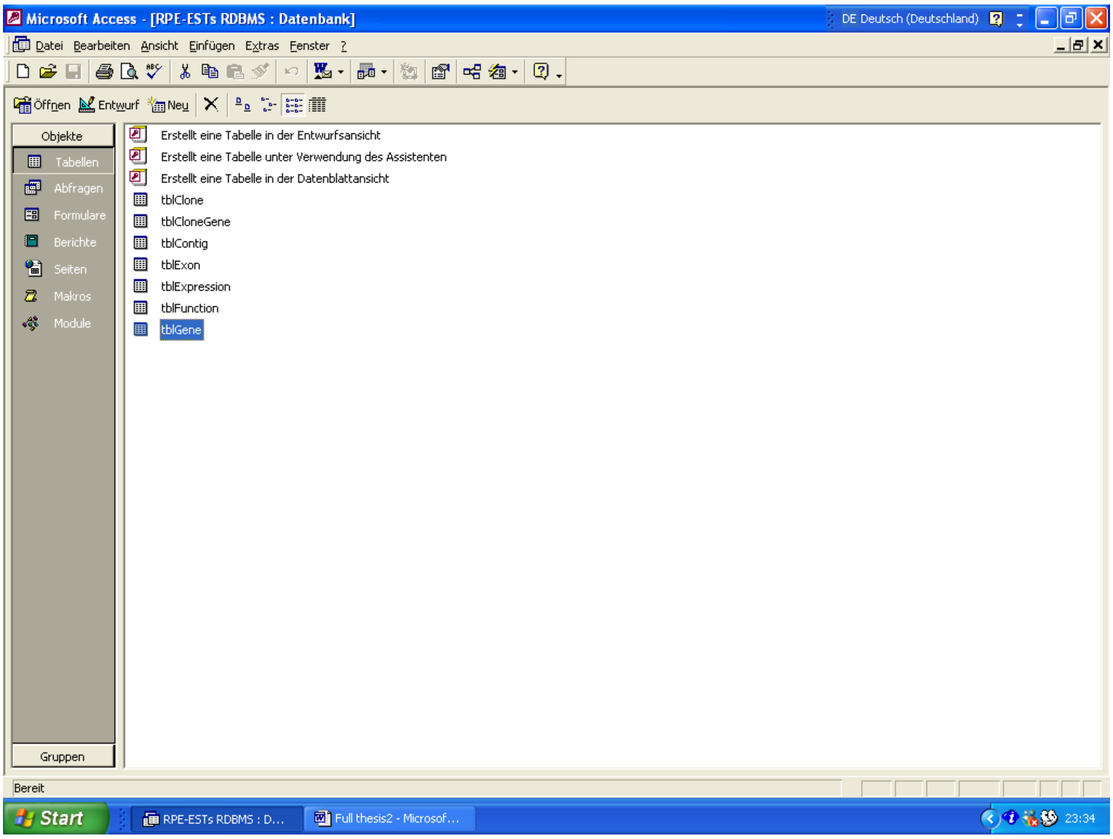
Gene ID
Class function
Subclass function
Protein motif
Protein domain
Biology of the gene

Exon ID
Gene ID
Start of Exon
End of Exon
Length of Exon
Exon Size (bp)
Nucleotide sequence of Exon
Amino Acid sequence of Exon
Exon F-primer
Exon R-primer

Clone name
Clone organism
Clone tissue
Clone source
Library type
Clone sequence
Clone length
PCR band Size (bp)

Contig #
**Category ID*
Contig consensus sequence
Contig length

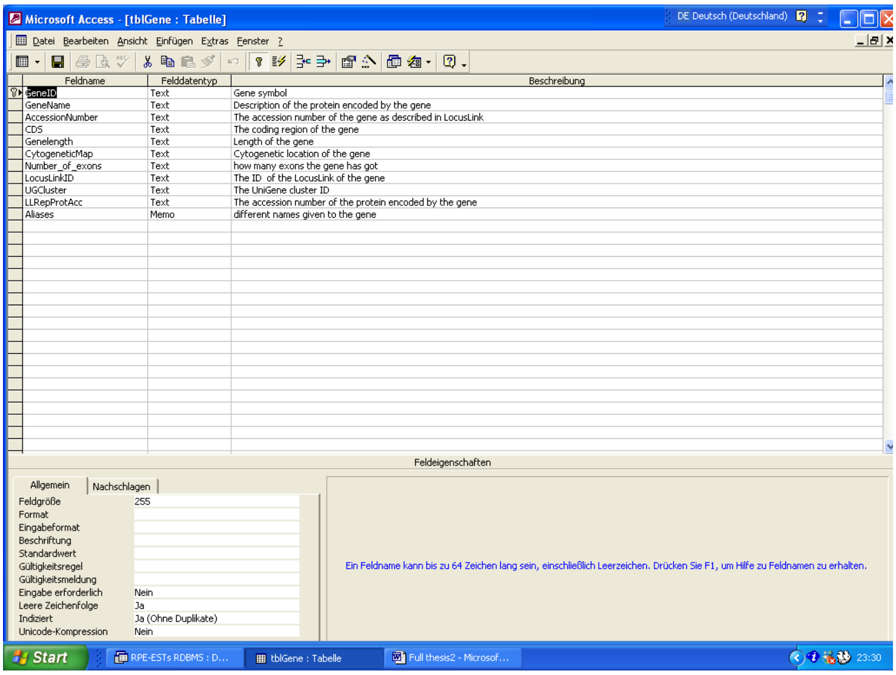
Clone name
Gene ID
**Contig #*
**Category ID*
% of homology
Place of homology – Query
Place of homology – Subject
Similarity DNA Source
Similarity with respect to CDS
Genomic Contig
Human Genomic Similarity
Contig_Percentage of homology
Contig_Place of Homology – Query
Contig_Place of Homology – Subject
Contig_commnets/Remarks
Subcategory
Chromosome ID
Cytogenetic map
Comments/Remarks



Erstellt eine Tabelle in der Entwurfsansicht
Erstellt eine Tabelle unter Verwendung des Assistenten
Erstellt eine Tabelle in der Datenblattansicht
tblClone
tblCloneGene
tblContig
tblExon
tblExpression
tblFunction
tblGene

Figure 6 The **tables** that were designed and constructed for the RDBMS. Each table represents an entity as summarized in Table 5.

An important task after creating tables, is to decide on the appropriate field type values, default values, captions, whether the field needs to be unique, whether the field needs to be indexed, any input masks and visual formatting. Figure 7 shows the fields created for the *GENE* table associated with the current application. A final important area involved with table creation is indexes. An index is used in a table much like index of a book. A type of index that every table should have is a primary key. Primary keys ensure uniqueness within a table and are vital for creating one-to-many relationships. Figure 8 shows the process of creating an index for *Gene ID* field in *GENE* table. In this case, *Gene ID* serves as the primary key for this table.

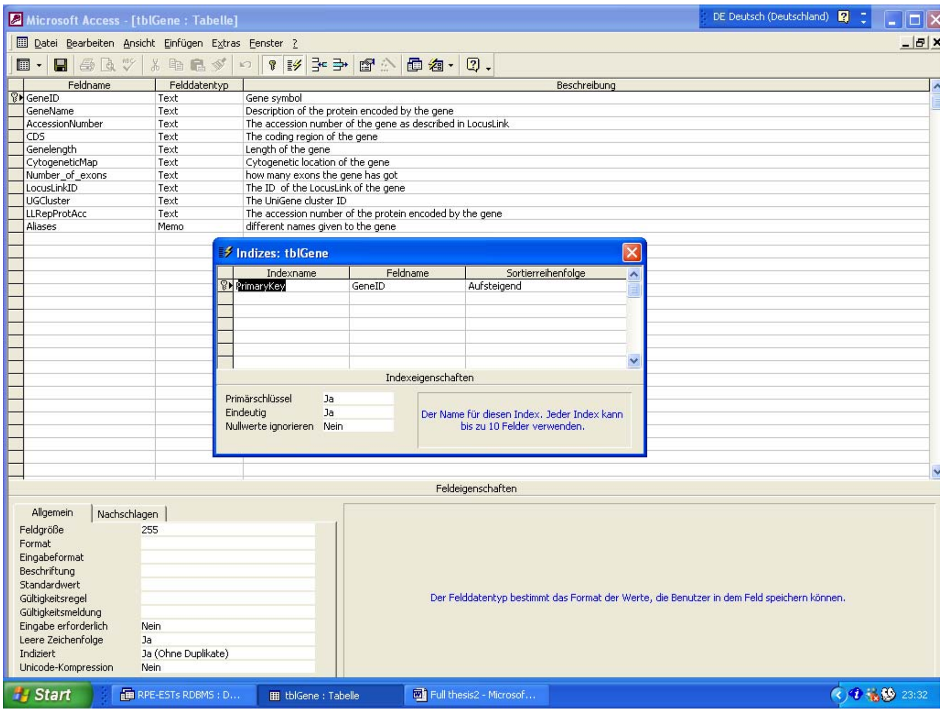

Microsoft Access - [tblGene : Tabelle]
DE Deutsch (Deutschland)
Datei Bearbeiten Ansicht Einfügen Extras Fenster ?
Feldeigenschaften
Allgemein
Nachschlagen
Start
RPE-ESTs RDBMS : D...
tblGene : Tabelle
Full thesis2 - Microsof...
23:30

Feldname	Felddatentyp	Beschreibung
GeneID	Text	Gene symbol
GeneName	Text	Description of the protein encoded by the gene
AccessionNumber	Text	The accession number of the gene as described in LocusLink
CDS	Text	The coding region of the gene
Genelength	Text	Length of the gene
CytogeneticMap	Text	Cytogenetic location of the gene
Number_of_exons	Text	how many exons the gene has got
LocusLinkID	Text	The ID of the LocusLink of the gene
UGCluster	Text	The UniGene cluster ID
LLRepProtAcc	Text	The accession number of the protein encoded by the gene
Aliases	Memo	different names given to the gene

Feldgröße	255
Format	
Eingabeformat	
Beschriftung	
Standardwert	
Gültigkeitsregel	
Gültigkeitsmeldung	
Eingabe erforderlich	Nein
Leere Zeichenfolge	Ja
Indiziert	Ja (Ohne Duplikate)
Unicode-Kompression	Nein

Ein Feldname kann bis zu 64 Zeichen lang sein, einschließlich Leerzeichen. Drücken Sie F1, um Hilfe zu Feldnamen zu erhalten.

Figure 7 The **GENE table** (tblGene) showing the fields created for the RDBMS. For example Gene Name, Gene ID, accession number. Each field represents an attribute for entity as summarized in Table 5 and in the third normal form (TNF) Table 7.



Feldname	Felddatentyp	Beschreibung
GeneID	Text	Gene symbol
GeneName	Text	Description of the protein encoded by the gene
AccessionNumber	Text	The accession number of the gene as described in LocusLink
CDS	Text	The coding region of the gene
Genelength	Text	Length of the gene
CytogeneticMap	Text	Cytogenetic location of the gene
Number_of_exons	Text	how many exons the gene has got
LocusLinkID	Text	The ID of the LocusLink of the gene
UGCluster	Text	The UniGene cluster ID
LLRepProtAcc	Text	The accession number of the protein encoded by the gene
Aliases	Memo	different names given to the gene

Indexname	Feldname	Sortierreihenfolge
PrimaryKey	GeneID	Aufsteigend

Primärschlüssel	Ja
Eindeutig	Ja
Nullwerte ignorieren	Nein

Der Name für diesen Index. Jeder Index kann bis zu 10 Felder verwenden.

Feldgröße	255
Format	
Eingabeformat	
Beschriftung	
Standardwert	
Gültigkeitsregel	
Gültigkeitsmeldung	
Eingabe erforderlich	Nein
Leere Zeichenfolge	Ja
Indiziert	Ja (Ohne Duplikate)
Unicode-Kompression	Nein

Der Felddatentyp bestimmt das Format der Werte, die Benutzer in dem Feld speichern können.

Figure 8 Primary key (index) created for the Gene table (tblGene). In this instance, *Gene ID* serves as the primary key for this table.

The RDBMS was designed and implemented into eight main tab sections or elements for easy access to information about clones, identified genes and their related information. Those eight tab components comprise the main menu of the system, as can be seen in Figure 9. The main menu includes: main, clone informaion, similarity/identity information, expression, function, known genes, search and others. The main tab component is containing Internet links to other valuable bioinformatic tools necessary for sequence analysis and annotations (Figure 9). Navigation buttons are always at the bottom, and users can exit from the system at any time.

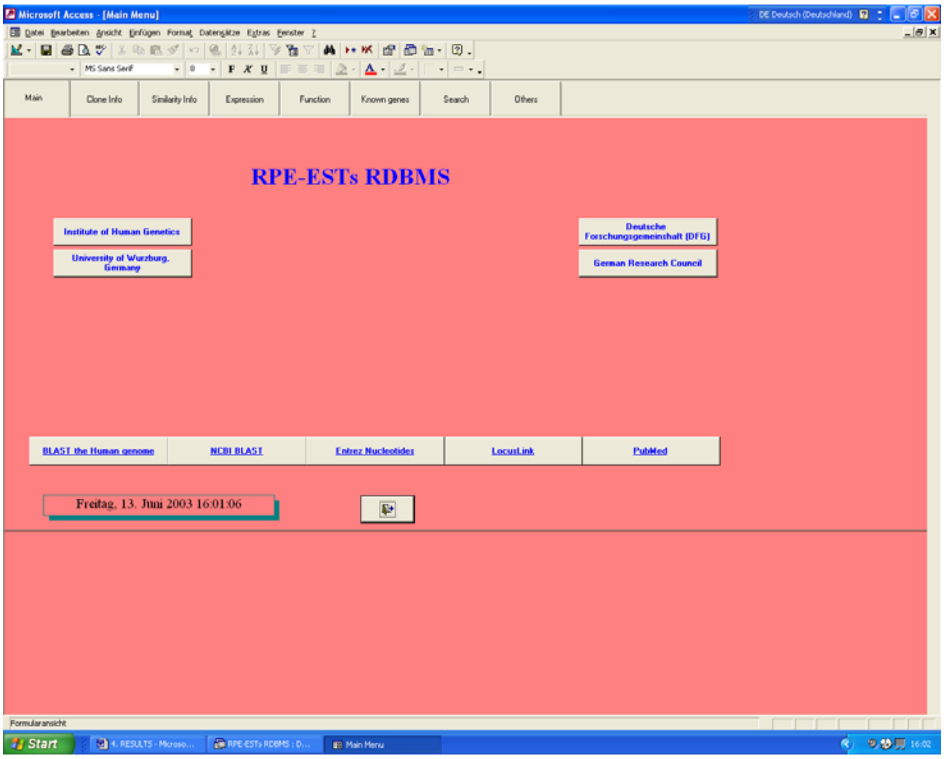
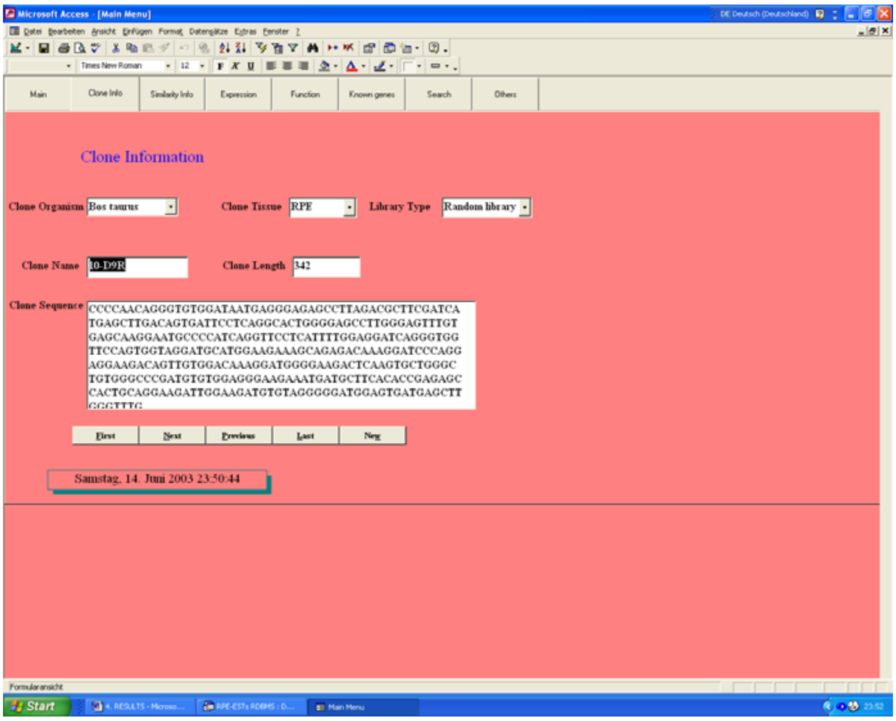


Figure 9 The **Main menu** of the RDBMS showing the various tabs representing forms that users can interact with to access the relevant information about clones, genes, expression, and functional profiling.

Figure 10 shows the fields necessary to capture most of the information needed about the clone in question. As described in Table 2, we need to keep information about RPE clones derived from the subtracted cDNA library, for example, clone name, clone length, clone sequence, clone tissue and clone organism. The first number in the Clone name refers to the RPE plate number, the letter and number that follows represent the plate co-ordinates. Using the clone sequence, Blast searches can be initiated and the result of the identity/similarity of the clone is entered into the similarity information page, just next to the clone page (Figure 11). Additionally, we need to keep track of the clone identity/similarity with reference to known, predicted gene ID, gene name, accession number, genomic contig, coding regions (CDS), Locus link ID and cytogentic map. We are also interested in clones that show no homology.

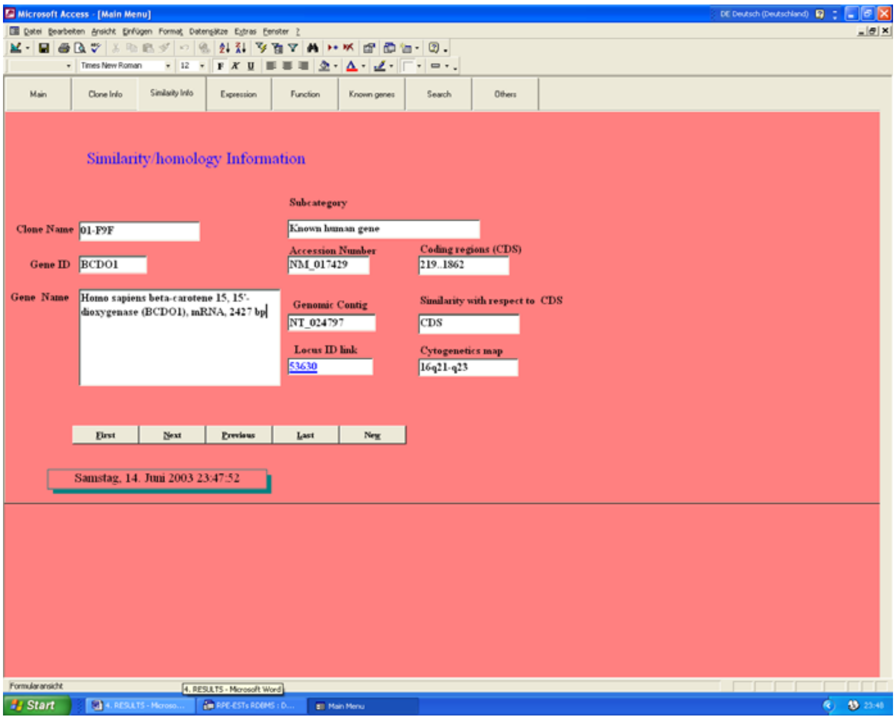


Clone Information

Clone Organism: Bos taurus
Clone Tissue: RPE
Library Type: Random library
Clone Name: 10-D9R
Clone Length: 342
Clone Sequence: CCCCAACAGGGTGTGGATAATGAGGGAGAGCCTTAGACGCTTCGATCATGAGCTTGACAGTGATTCCTCAGGCACTGGGGAGCCTTGGGAGTTTGTGAGCAAGGAATGCCCCATCAGGTTCCTCATTTTGGAGGATCAGGGTGGTTCCAGTGGTAGGATGCATGGAAGAAAGCAGAGACAAAGGATCCCAGGAGGAAGACAGTTGTGGACAAAGGATGGGGAAGACTCAAGTGCTGGGCTGTGGGCCCGATGTGTGGAGGGAAGAAATGATGCTTCACACCGAGAGCCACTGCAGGAAGATTGGAAGATGTGTAGGGGGATGGAGTGATGAGCTTGGGTTTG

Samstag, 14. Juni 2003 23:50:44

Figure 10 The **Clone page** showing the fields necessary to capture most of the information needed about the clone in question. The first number in the Clone name refers to the RPE plate number, the letter and number that follow representing the plate co-ordinates.



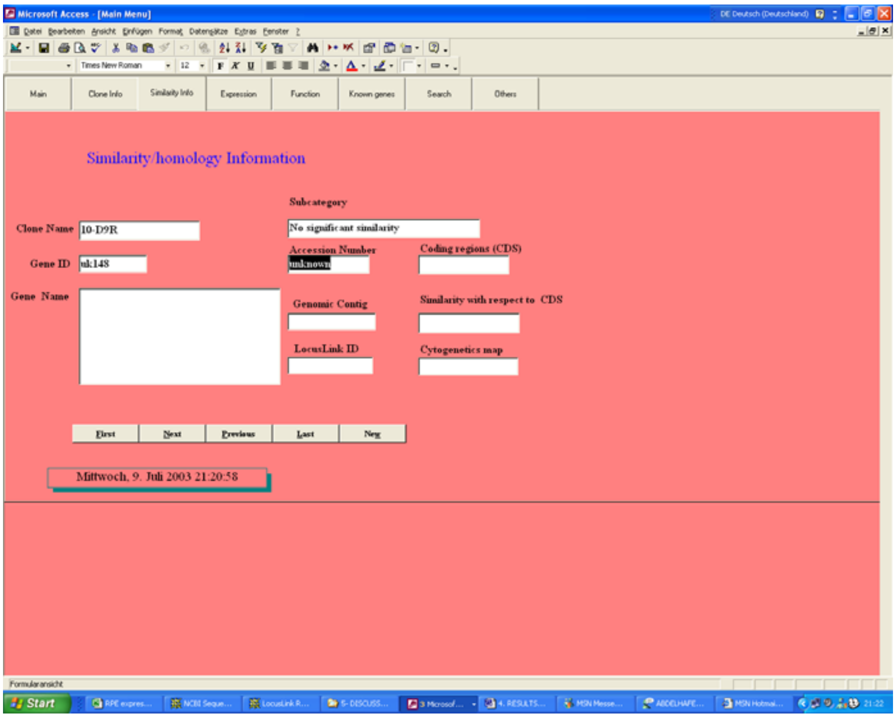
Similarity/homology Information

Clone Name: 01-F9F
Gene ID: BCDO1
Gene Name: Homo sapiens beta-carotene 15, 15'-dioxygenase (BCDO1), mRNA, 2427 bp
Subcategory: Known human gene
Accession Number: NM_017429
Coding regions (CDS): 219..1862
Genomic Contig: NT_024797
Similarity with respect to CDS: CDS
Locus ID link: 53630
Cytogenetics map: 16q21-q23

Samstag, 14. Juni 2003 23:47:52

Figure 11 The **Similarity tab page** showing the clone 01-F9F with similarity to BCDO1 gene on chromosome 16 band q21-q23 with GenBank acc. No. NM_017429.

In this case, the laboratory-intern gene symbol can easily be entered in the Gene ID field, and the “no homology” status can be described in the subcategory field, either as “unknown with or without exon-intron boundaries” or “no significant similarity was found” (Figure 12).



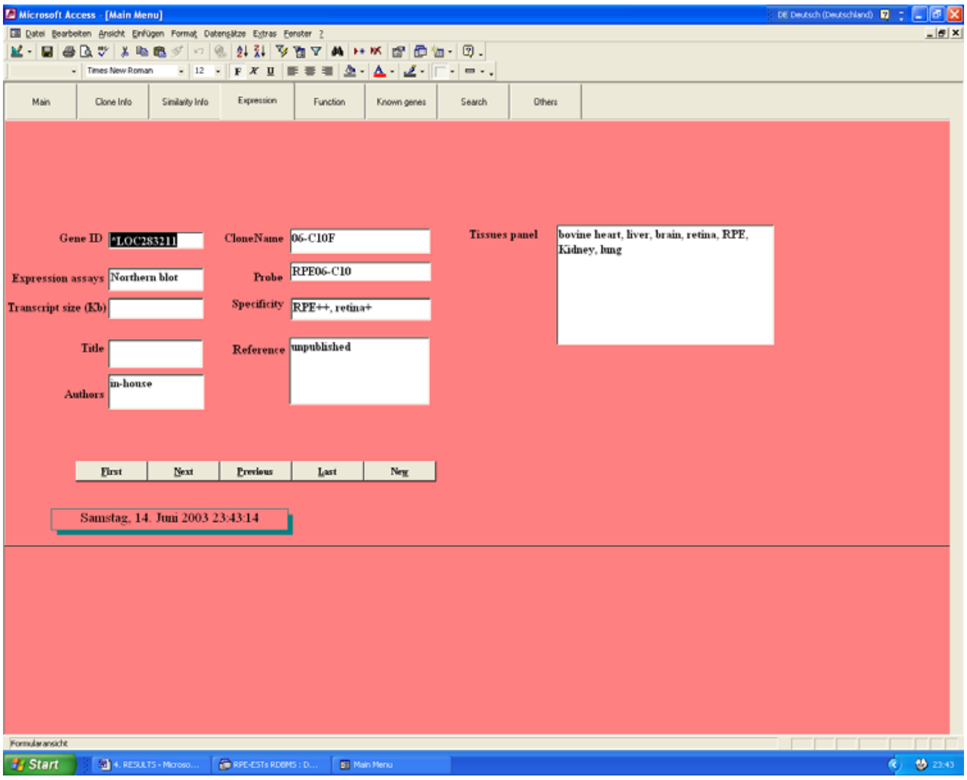
Similarity/homology Information

Clone Name: 10-D9R
Gene ID: uk148
Gene Name:
Subcategory: No significant similarity
Accession Number: unknown
Coding regions (CDS):
Genomic Contig:
Similarity with respect to CDS:
LocusLink ID:
Cytogenetics map:

Mittwoch, 9. Juli 2003 21:20:58

Figure 12 The **Similarity tab page** showing the clone 10-D9R that showed no significant similarity as denoted in subcatergy field. The unknown status of this clone is shown as the laboratory-intern gene symbol “uk148” in Gene ID field and “unknown” in the accession number field.

Another area of interest which requires the storage of information is the expression profiling of either the clones or identified genes. Genes expressed in a certain tissue can be assumed to be important for the function of this tissue. The expression data can have two sources; either from the published literature mainly for known and predicted genes or from experiments using Northern blot or RT-PCR for unknown transcripts or for known genes that have no expression data in the published literature. Figure 13 shows the clone RPE06-C10 that was found using Northern blot analysis on bovine RNA samples to be expressed in retina and RPE with abundant expression in RPE. No expression was detected in other tissues examined. Analysis of the complete sequences of SSRPE cDNA library has identified a number of many known and predicted genes. With complete coding sequences, protein sequences can be examined for motifs, domains, and biochemical characteristics that may suggest function. The most challenging problem will then be to determine the functions of these genes. Figure 14 shows the form that was designed and implemented to capture, and to display, the most important information needed about the function of the known and predicted genes. A variety of tools, including BLASTN, BLASTX, LocusLink, SwissProt, Source, Pfam and published literature were used for the functional annotation of the known and predicted genes. For example, in Figure 14 the **Function tab page** shows the class function and subclass function of the *BCDO1* gene as vitamin A metabolism/transport and vitamin A metabolism respectively.



Gene ID *LOC283211

CloneName 06-C10F

Tissues panel bovine heart, liver, brain, retina, RPE, Kidney, lung

Expression assays Northern blot

Probe RPE06-C10

Transcript size (Kb)

Specificity RPE++, retina+

Title

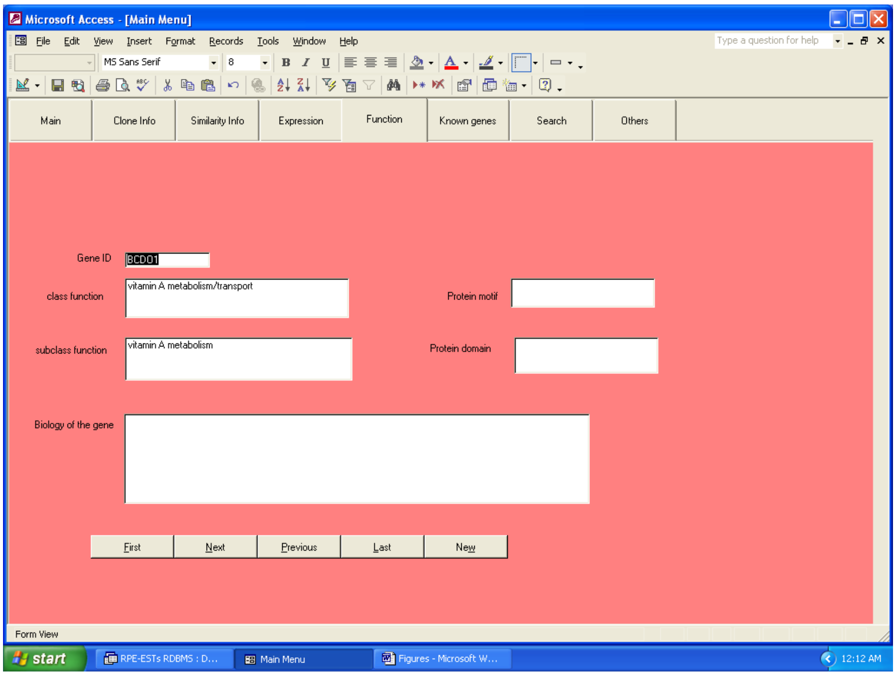
Reference unpublished

Authors in-house

Samstag, 14. Juni 2003 23:43:14

Figure 13 The **Expression page** showing the clone RPE06-C10 that was found, using Northern blot hybridization on bovine RNA, to be expressed in retina and RPE with abundant expression in RPE. No expression was detected in other tissues examined.

Efforts were made to normalize the redundancy of the known genes and to group them into clusters that represent transcripts from the same gene (Figure 5A). Figure 15 shows the information restricted to known genes only and its related information such as expression and function as well as a search option to find the desired gene. Figure 16 shows other information such as clone ID, clone source, clone microtitre plate, blast date, and revision date. These are for monitoring and organisation purposes only. For example, it is extremly important to know the first date of blasting the clone in question as well as the last revision date before any further decision is taken. In addition, clone source is for statistic purposes to know if the clone was sequenced in-house or at Lynkeus Biotech.



Gene ID: BCDO1

class function: vitamin A metabolism/transport

subclass function: vitamin A metabolism

Figure 14 The **Function tab page** showing the class function and subclass function of the *BCDO1* gene.

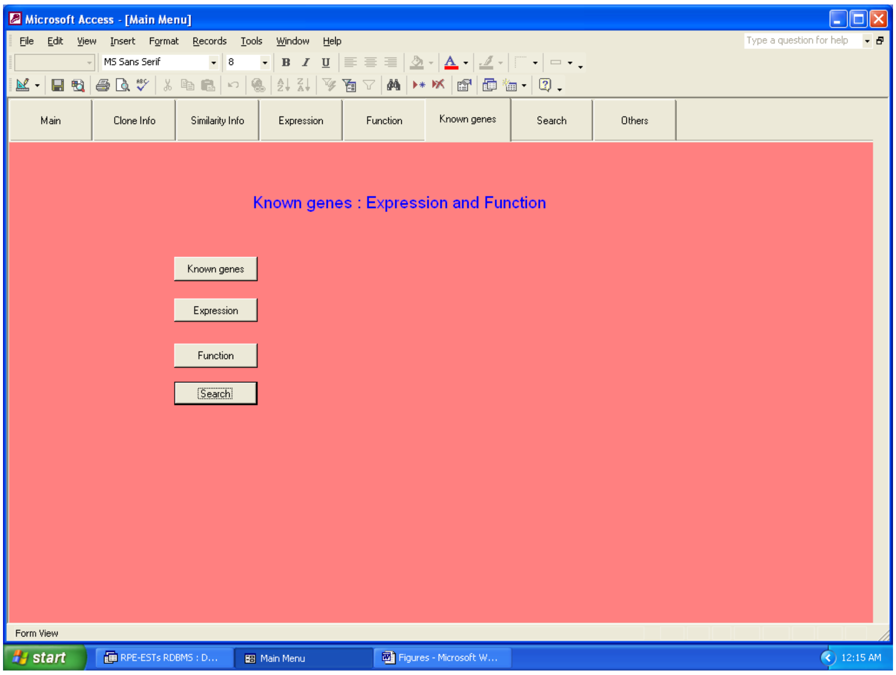
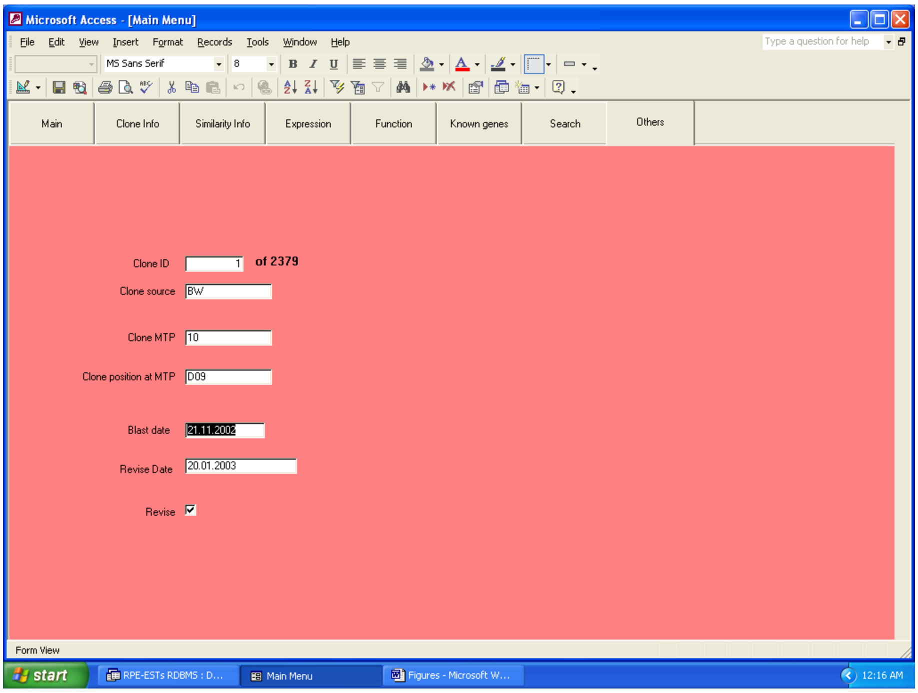


Figure 15 The **Known genes tab page** showing the information restricted to known genes only as well as a search option to find a certain gene and its related information such as expression and function.

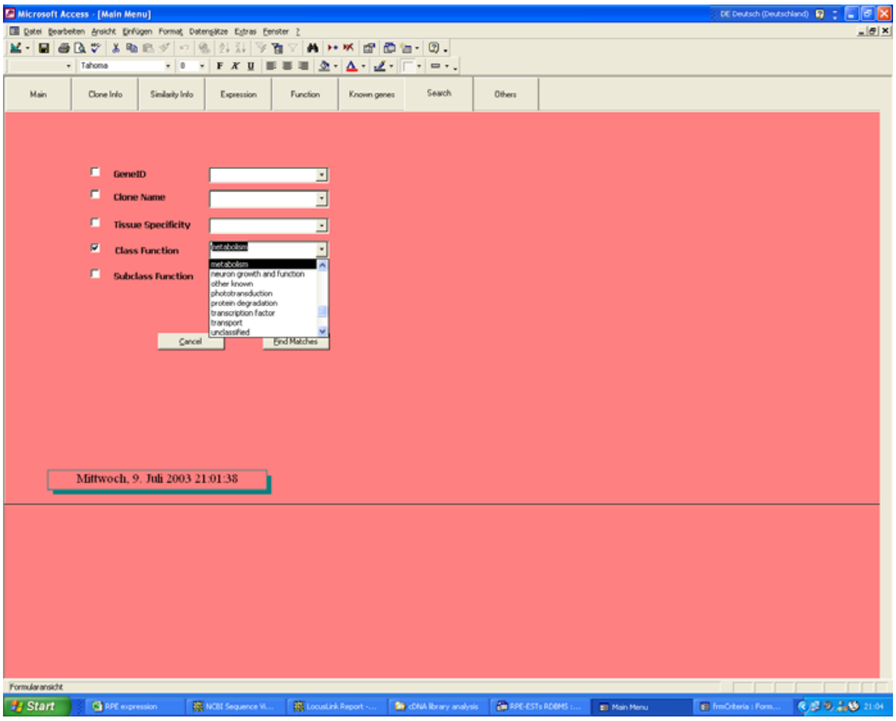


Clone ID 1 of 2379
Clone source BW
Clone MTP 10
Clone position at MTP D09
Blast date 21.11.2002
Revise Date 20.01.2003
Revise ☑

Figure 16 The **others tab page** showing other information such as clone ID, clone source, clone microtitre plate, blast date, and revise date. These are for monitoring and organisation purposes only.

The **Search form** Query by Form (QBF) interface contains text boxes or combo boxes for each of the criteria that the user of the RDBMS can specify (Figure 17). The user clicks the check box to indicate that a particular criterion to be used. The process of clicking this check box enables the text boxes or combo boxes in which the values for the criterion are then entered. The screenshot on Figure 17 illustrates how the search form could be used to ask the question: **Show me all the RPE known genes that function in metabolic pathways?** Figure 18 shows the results of such a search.

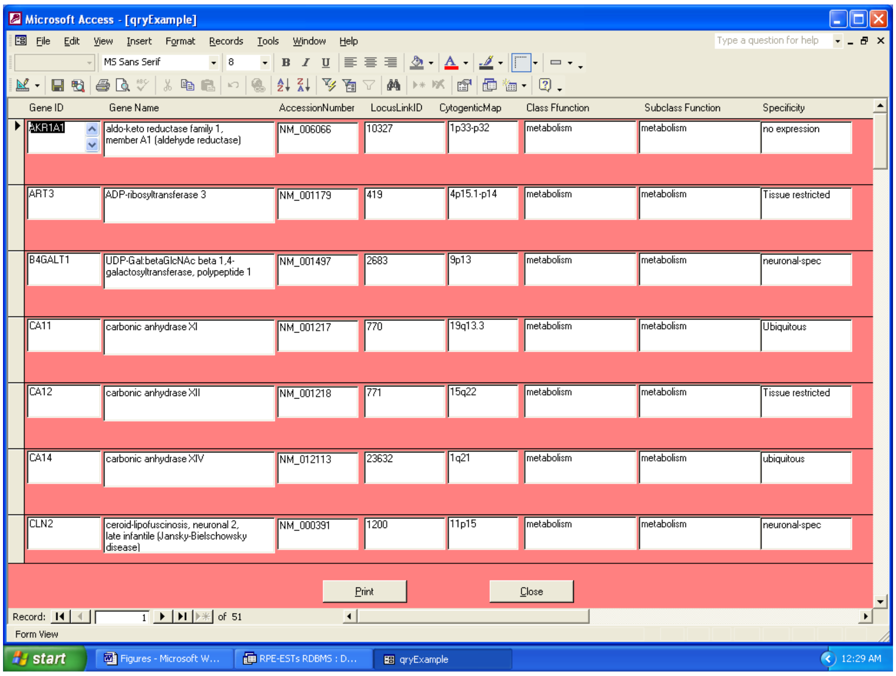
When the user activates the *Find Matches* button, the results form (Figure 17) is displayed containing the records that match the criteria specified by the user (Figure 18). When the *Print* button is pressed, the report containing those records is previewed on screen (Figure 19).



☐ GeneID
☐ Clone Name
☐ Tissue Specificity
☑ Class Function: metabolism
☐ Subclass Function

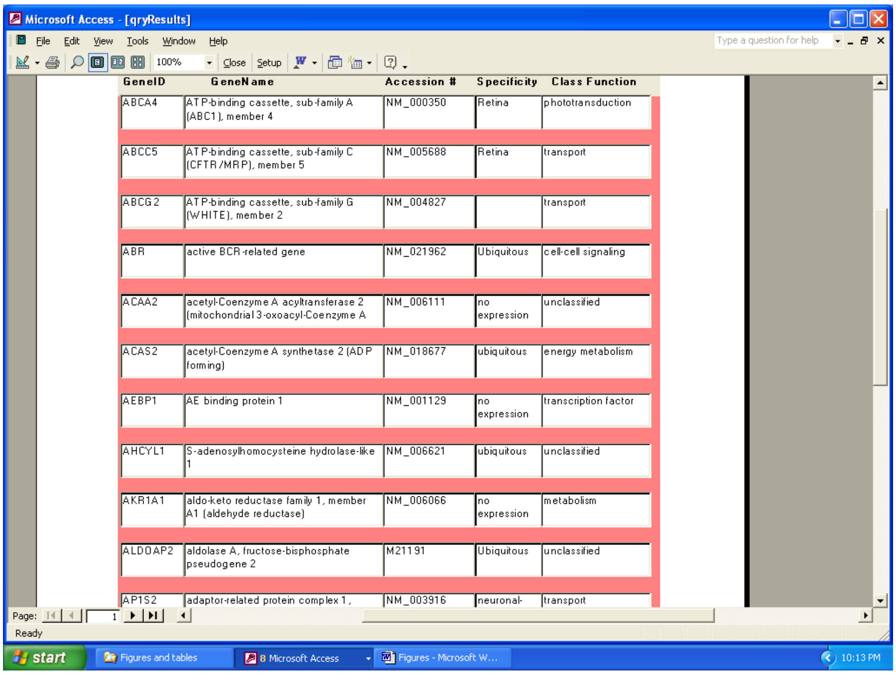
Mittwoch, 9. Juli 2003 21:01:38

Figure 17 Search form showing Query by Form (QBF) interface illustrates how the search form could be used to ask the question: **Show me all the RPE known genes that function in metabolic pathways?** Then, as the Find Matches is activated, the result form should display the records that match the criteria specified by the user (Figure 18).



Gene ID	Gene Name	AccessionNumber	LocusLinkID	CytogenticMap	Class Ffunction	Subclass Function	Specificity
AKR1A1	aldo-keto reductase family 1, member A1 (aldehyde reductase)	NM_006066	10327	1p33-p32	metabolism	metabolism	no expression
ART3	ADP-ribosyltransferase 3	NM_001179	419	4p15.1-p14	metabolism	metabolism	Tissue restricted
B4GALT1	UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 1	NM_001497	2683	9p13	metabolism	metabolism	neuronal-spec
CA11	carbonic anhydrase XI	NM_001217	770	19q13.3	metabolism	metabolism	Ubiquitous
CA12	carbonic anhydrase XII	NM_001218	771	15q22	metabolism	metabolism	Tissue restricted
CA14	carbonic anhydrase XIV	NM_012113	23632	1q21	metabolism	metabolism	ubiquitous
CLN2	ceroid-lipofuscinosis, neuronal 2, late infantile (Jansky-Bielschowsky disease)	NM_000391	1200	11p15	metabolism	metabolism	neuronal-spec

Figure 18 Results form showing RPE known genes that function in metabolic pathways, as well as their *in silico* expression profiling, cytogenetic map locations, LocusLink IDs, and accession numbers.



GeneID	GeneName	Accession #	Specificity	Class Function
ABCA4	ATP-binding cassette, sub-family A (ABC1), member 4	NM_000350	Retina	phototransduction
ABCC5	ATP-binding cassette, sub-family C (CFTR/MRP), member 5	NM_005688	Retina	transport
ABCG2	ATP-binding cassette, sub-family G (WHITE), member 2	NM_004827		transport
ABR	active BCR-related gene	NM_021962	Ubiquitous	cell-cell signaling
ACAA2	acetyl-Coenzyme A acyltransferase 2 (mitochondrial 3-oxoacyl-Coenzyme A	NM_006111	no expression	unclassified
ACAS2	acetyl-Coenzyme A synthetase 2 (ADP forming)	NM_018677	ubiquitous	energy metabolism
AEBP1	AE binding protein 1	NM_001129	no expression	transcription factor
AHCYL1	S-adenosylhomocysteine hydrolase-like 1	NM_006621	ubiquitous	unclassified
AKR1A1	aldo-keto reductase family 1, member A1 (aldehyde reductase)	NM_006066	no expression	metabolism
ALDOAP2	aldolase A, fructose-bisphosphate pseudogene 2	M21191	Ubiquitous	unclassified
AP1S2	adaptor-related protein complex 1,	NM_003916	neuronal-	transport

Figure 19 Report form showing an example of RPE known genes

2. Analysis of the SSRPE cDNA library

In the present study, 2379 expressed sequence tags (ESTs) from SSRPE cDNA library of bovine were analyzed. The length of inserts was in the range of 0.2 to 2 kb. The average readable sequence length, on which the following analysis was based, was approximately 300-600-bp. To overcome the problem of EST redundancy and to increase the length of the sequences, facilitating annotation by homology searches, a clustering process was performed. In this clustering process, RPE EST sequences that have a sufficient region of similarity are joined into a cluster. Thus, sequences possessing overlapping regions and representing a single gene are joined into the same cluster, therefore decreasing redundancy.

2.1 Phase I: Analysis of 1002 ESTs

In a first phase, 1002 SSRPE cDNA clones were sequenced and blasted to available sequences in the non-redundant GenBank/EMBL and dbEST databases (January – May 2001). Our analysis revealed that a total of 590 ESTs (59%) represent known or predicted genes, while 395 ESTs (39%) did not match to known transcripts. Of these 245 ESTs matched to ESTs and genomic sequences, and 150 ESTs showed no homology to sequences available in the nucleotide databases (Table 8).

Correction for redundancy in the category of ESTs representing known genes was done by cluster analysis and sequence alignments to known transcripts, and showed that on average a single gene was represented by almost 5 ESTs (factor 4.8) (Table 8). On the other hand, the redundancy in the remaining categories was lower as many of the non-overlapping ESTs corresponding to only partially known genes could not be linked together. As a result of redundancy correction, 379 normalized transcripts were obtained (Table 8).

Table 8 Summary of 1002 ESTs derived from RPE cDNA library (as of May 2001)

		Clones		**Unique transcripts**	
	Category	**n**	**%**	**n**	**%**
I	Known genes	580	57.9	120	31.7
II	predicted genes	10	1.0	7	1.8
III	ESTs/genomic sequences	245	24.5	163	43.0
IV	No homology	150	15.0	86	22.7
V	Mitochondrial transcripts	17	1.6	3	0.8
	Total	1002		379	

The representative clone sequences of the 376 in silico normalized transcripts (minus the 3 mitochondrial transcripts) were blasted against the human genome draft sequence (http://www.ncbi.nlm.nih.gov/genome/seq/page.cgi?F=HsBlast.html&&ORG=Hs). Many of the singleton ESTs not previously linked to a specific mRNA transcript could be located within the immediate 3’ UTR (untranslated region) of known or predicted human genes. A second round of normalization (as of October 2001) resulted in the identification of 168 known, 51 predicted, and 1 unknown transcript with exon-intron boundaries. In addition, 14 transcripts without exon-intron boundaries, and 41 ESTs showed no significant similarity (Figure 20).

In summary, a total of 1002 bovine RPE-ESTs represent 168 known human genes and approximately 107 predicted or as yet unknown genes. To evaluate the SSRPE cDNA library for the degree of enrichment for RPE-specific transcripts we collected *in silico* data on expression profiling of the 168 known genes from published sources. Overall, data were obtained for 153 genes. Approximately 17% (26/153) of these genes were RPE and/or retina-specific and approximately 5% (7/153) were specifically expressed in retina and brain. Extrapolating these data leads us to predict that the approximately 107 unknown genes identified in the first phase of this study could include 15-20 novel RPE/Retina-specific genes.

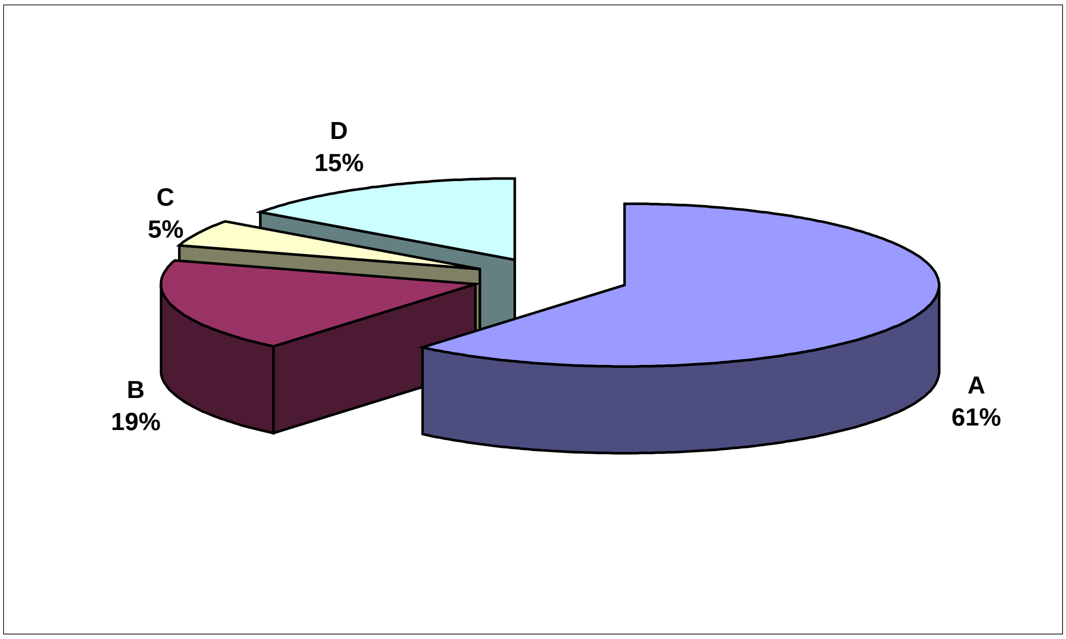


Figure 20 Results of blast searches of 376 normalized transcripts (as of October 2001) 376 normalized cDNA clones (minus the 3 mitochondrial transcripts) were compared to the human genome draft sequence. The clones that showed homology are indicated as percentages and are shown at the bottom of the diagram. **A**. Identities to known human genes. **B**. Identities to human predicted genes. **C.** Similarity to unknown transcripts **D**. No significant similarity to human sequences.

2.2 Phase II: Analysis of additional 1377 ESTs

In a second phase, additional 1377 SSRPE cDNA clones were sequenced and blasted against a recent release of the human genome draft sequence. This raised the total number of the bovine RPE cDNA clones to 2379. Contig assembly program CAP3 (Huang and Madan, 1999) was used to assemble the high quality sequences into clusters or singletons. 1.2% of the 2379 RPE-ESTs contains vector sequences and was excluded from further analysis. 5% of the RPE-ESTs showed homology to multipe chromosomes and were not included in the further assembly process. The rest of the ESTs (2245) were assembled into 175 contigs and 509 singletons, which revealed 684 unique genes in the data set. The frequency of EST distribution after CAP3 analysis and BLAST searches (as of March 2003) is shown in Table 9. Most clusters constained 1-3 ESTs and the largest cluster contained 151 ESTs.

Table 9 Summary of the abundance occurring in the 2379 RPE-ESTs*(as of March 2003)

Frequency of EST hits	Known	Uknown	Gene Names	Accession Number
151	1	-	VMD2	NM_004183
87	1	-	TTR	NM_000371
79	1	-	TRPM3	XM_036123
78	1	-	TNRC15	XM_209467
68	1	-	CDH3	NM_001793
63	1	-	RLBP1	XM_007697
52	1	-	RPE65	NM_000329
44	1	-	GSTM	NM_000561, NM_000850
39	-	1	-	-
37	2	-	RHO, RGR	NM_000539, NM_002921
36	1	-	SERPINF1	NM_002615
35	3	-	RDH10	NM_172037
			DKFZP564K1964	NM_015544
			SLC2A1	NM_006516
33	1	-	LRAT	NM_004744
32	1	-	LOC283211	XM_210938
28	-	1	-	-
27	1	-	CA14	NM_012113
26	-	1	-	-
22	1	-	SLC4A5	NM_021196
19	1	-	CST3	NM_001322
18	3	-	RCV1, MAOB	NM_002903, NM_000898
			FMOD	NM_002023
17	1	-	ATP1B2	NM_001678
15	2	-	DKFZP761G0122	NM_152661
			SKIP	NM_016532
14	-	1	-	-
13	2	-	BCDO1, TRPM1	NM_017429, NM_002420
12	1	-	GPM6B	NM_005278
11	2	-	ABHD6	NM_020676
			LOC119587	XM_058409
10	1	-	RDS	NM_000322
9	1	-	KIAA1576	NM_020927
8	5	2		
7	5	-		
6	5	2		
5	16	-		
4	9	2		
3	30	5		
2	47	11		
1	192	317		
Total	**341**	**343**		

*The 2379 RPE ESTs were sequenced in two phases. In phase I, 1002 ESTs were sequenced in-house, whereas in phase II, 1377 were sequenced at Lynkus Company (Würzburg).

The analysis of 2245 ESTs revealed that some of the RPE genes were represented in the library by more than one EST. The top most abundant ESTs are shown in Table 9. From the 684 unique transcripts identified in this study, 341 (49.2%) were attributed to known genes (Table 9) according to NCBI database using the BLAST program (Altschul et al., 1990) and the human genome draft sequence (http://www.ncbi.nlm.nih.gov/genome/ seq/page.cgi?F= HsBlast .html&&ORG=Hs). Consequently, no hits were found for 343 (50.2%) (Table 9). Many of these, if not all, should represent novel RPE genes, but further investigations will be required.

3. Functional profiling of the reported known RPE genes

A variety of tools, including BLASTN, BLASTX, LocusLink, SwissProt, Source, Pfam and data from published literature were used for the functional annotation of the known and predicted genes. The identified 341 known and predicted genes were found to represent 18 different functional groups (Figure 21). A comprehensive list of these functional groups including a list of all genes can be found in the Appendix (VIII, 1). Of the 341 known genes, no function could be assigned to 24.4%. Genes involved in metabolic pathways are represented and corresponded to 14.5% of the total number of annotated genes. 12.5% of the known genes are involved in various types of transport. We also found that 10.2% of the known genes are involved in cell-cell signalling. 5.8% of the known genes fall into the cell defence group as RPE is subjected to high oxidative stress. 2.6% of the identified known genes in this study are involved in vitmain A metabolism and transport, including RPE65, RLBP1, RBP1, LRAT, BCDO1, RBP4, and RDH5. 3.8% of the reported genes are involved in transcriptional factors, including OTX2 (orthodenticle homolog 2 (Drosophila), NRL (neural retina leucine zipper). 2% of the reported genes are involves in phototransduction and 1% are lysosomal enzymes.

4. Comparison to similar studies

Wistow et al. generated 10,000 ESTs from the human RPE/choroid for the NEIBank project (Wistow et al., 2002a). Comprehensive analysis of the NEIBank RPE unique gene clusters revealed that 3820 attributed to known genes in the public databases and the rest of 2480 have no matches (as of March 2003). Comprehensive comparison of the NEIBank 3820 unique RPE clusters with our datasets of 341 RPE known genes shows that only 146 are overlapping between the two libraries (Figure 23 and Table 10). 42.82% of our identified RPE genes are overlapping with only 3.82% of NEIBank RPE known genes. 10 of the 146 overlapping genes are involved in retinal dystrophy diseases (shown as bold and underlined in Table 10).

Crosstab query was designed and implemented to compare the 3820 NEIBank RPE/choroid and our 341 SSH-RPE known genes (Figure 22). The comparison shows that only 146 are overlapping between the two libraries (Figure 23 and Table 10). This illustrates the potential of MS access in importing external tables and compares them with the existing tables in the database using the crosstab query feature.

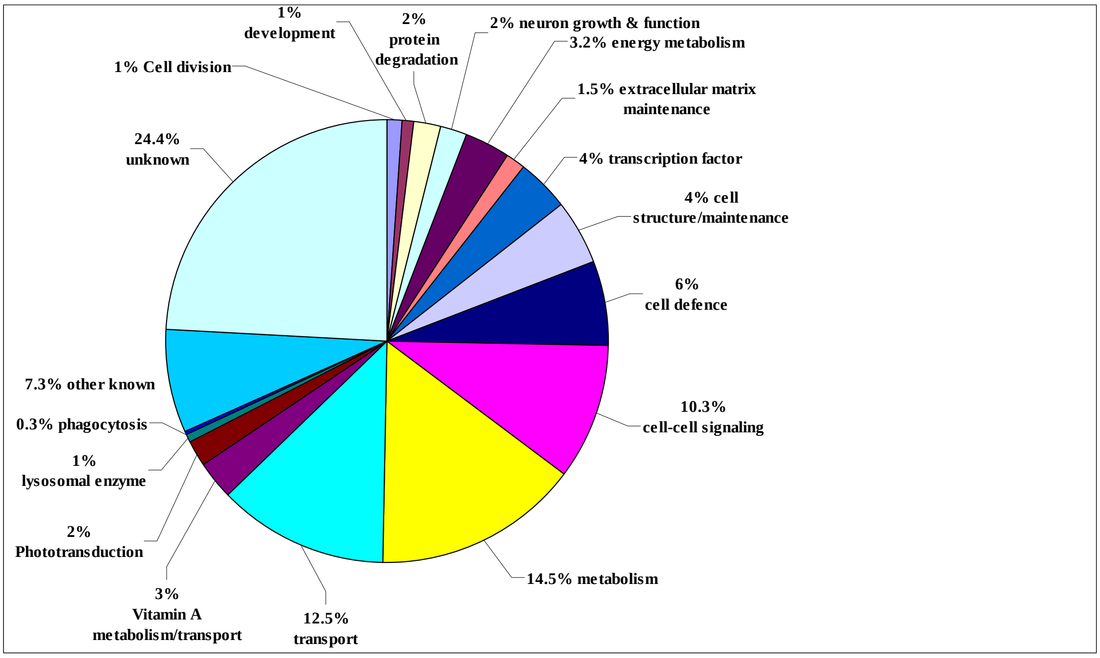
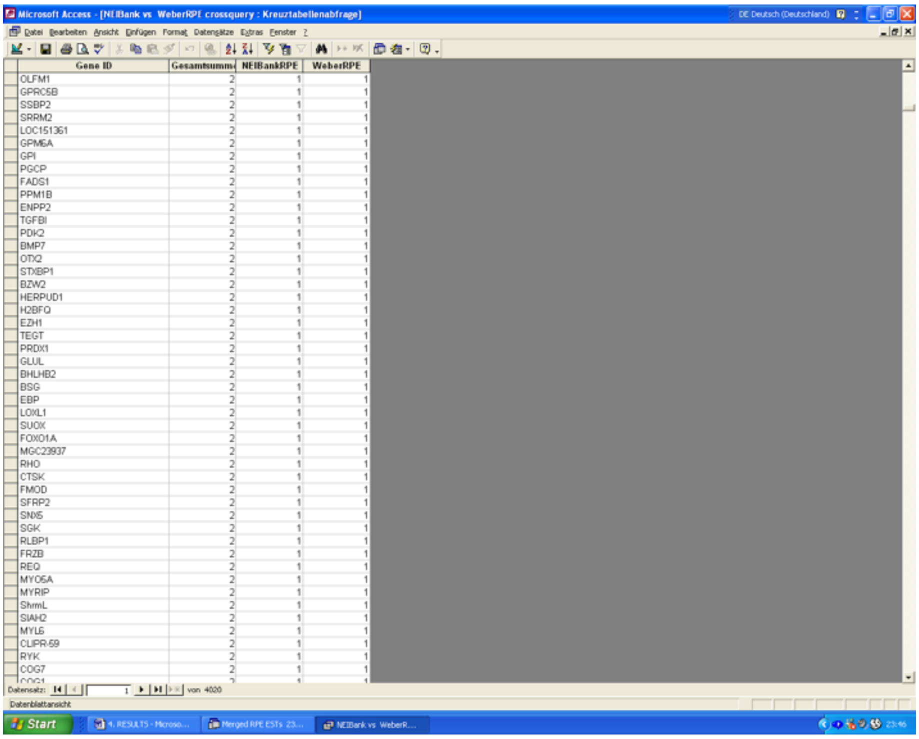


Figure 21 Functional groups of RPE known genes identified in this study. A total of 341 non-redundant genes were classified into 18 groups according to their probable function.



Gene ID	Gesamtsumme	NEIBankRPE	WeberRPE
OLFM1	2	1	1
GPRC5B	2	1	1
SSBP2	2	1	1
SRRM2	2	1	1
LOC151361	2	1	1
GPM6A	2	1	1
GPI	2	1	1
PGCP	2	1	1
FADS1	2	1	1
PPM1B	2	1	1
ENPP2	2	1	1
TGFBI	2	1	1
PDK2	2	1	1
BMP7	2	1	1
OTX2	2	1	1
STXBP1	2	1	1
BZW2	2	1	1
HERPUD1	2	1	1
H2BFQ	2	1	1
EZH1	2	1	1
TEGT	2	1	1
PRDX1	2	1	1
GLUL	2	1	1
BHLHB2	2	1	1
BSG	2	1	1
EBP	2	1	1
LOXL1	2	1	1
SUOX	2	1	1
FOXO1A	2	1	1
MGC23937	2	1	1
RHO	2	1	1
CTSK	2	1	1
FMOD	2	1	1
SFRP2	2	1	1
SNX6	2	1	1
SGK	2	1	1
RLBP1	2	1	1
FRZB	2	1	1
REQ	2	1	1
MYO5A	2	1	1
MYRIP	2	1	1
ShrmL	2	1	1
SIAH2	2	1	1
MYL6	2	1	1
CLIPR-59	2	1	1
RYK	2	1	1
COG7	2	1	1

Figure 22 Crosstab query used to compare the 3820 NEIBank RPE/choroid and the 341 our SSRPE known genes. The comprehensive comparison shows that only 146 are overlapping between the two libraries (Figure 23 and Table 10).

Table 10 The 146 overlapping RPE genes between NEIBank and SSH-RPE library, showing the gene symbols

ABCA4
ABR
ACAA2
AEBP1
AHCYL1
ARL6IP
ATP1A3
ATP1B2
ATP5B
ATP5G2
ATP5L
ATP6V0A1
B4GALT1
BHLHB2
BMP7
BSG
BZW2
CA12
CLIC6
CLIPR-59
CLU
COG1
COG7
COX4I1
CST3
CTSK
DDOST
DKFZp564D206
DKFZP564K1964
EBP
EFEMP1
ENPP2
ERH
EZH1
FADS1
FLJ10803
FLJ11305
FLJ12287
FLJ32069
FLOT2
FMOD
FOXO1A
FRZB
FTH1
GALNT10
GAPD
GLUL
GNAT1
GNB3
GOT1
GOT2
GPI
GPM6A
GPRC5B
GPX4
H2BFQ
HERPUD1
HNRPDL
HSPC111
HSPCB
IPO13
JWA
KIAA0157
KIAA0446
KIAA1576
LASS2
LDHB
LOC151361
LOXL1
MAGED1
MAP2K1
MAPKBP1
MFRP
MGC23937
MGC32043
MLF2
MYL6
MYO5A
MYRIP
NDUFA4
NICE-3
NIFIE14
NPR2
NUCB1
OLFM1
OTX2
PDK2
PGCP
PGRMC1
PITPN
PPM1B
PRDX1
PRKWNK4
PRSS11
RAB5B
RAB7
RDS
REQ
RGR
RHO
RLBP1
RPE65
RPS3
RYK
SAG
SAT
SCMH1
SDF2
SDHA
SERPINA5
SERPINF1
SFRP2
SGK
ShrmL
SIAH2
SLC13A3
SLC16A1
SLC22A8
SLC2A1
SLC4A5
SLC6A13
SLC9A3R1
SMPD1
SNX5
SPOCK
SPTAN1
SRRM2
SSBP2
STUB1
STXBP1
SUOX
TEGT
TGFBI
TIMP3
TNRC15
TTR
TRIM41
TSC22
TSPAN-
TUBG1
TYRP1
UBE1
VEGF
VMD2
WBP1
XT3

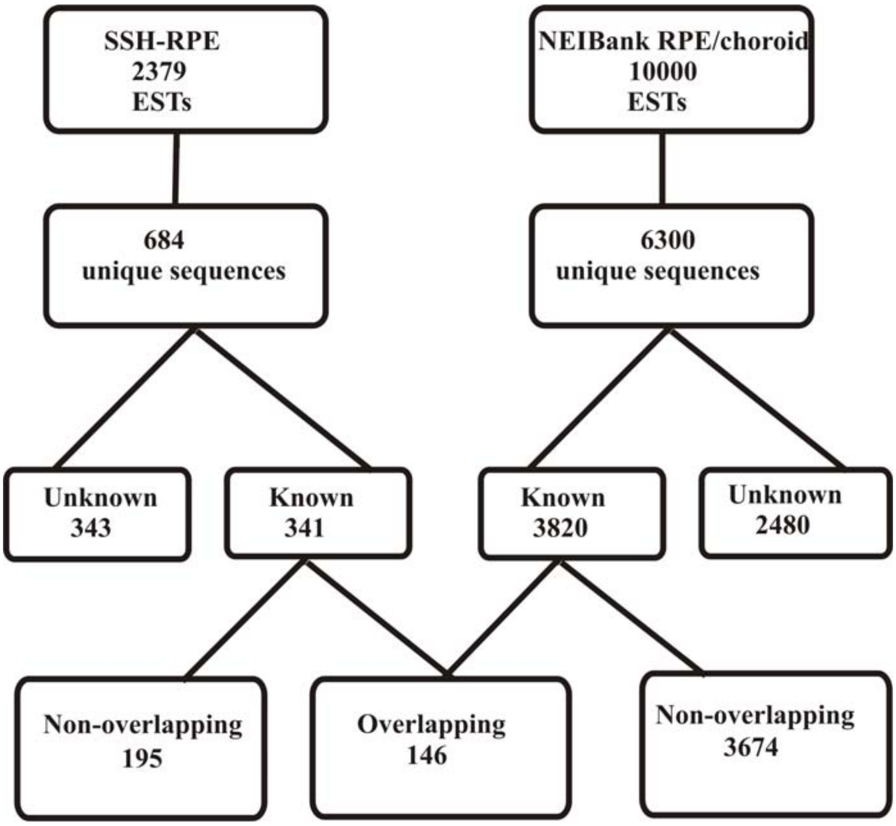


Figure 23 Breakdown of the ESTs sequences of NEIBank RPE/choroid and SSH-RPE library showing the 146 overlapping RPE genes between the two libraries.

5. Cloning and characterization of RPE01-D2 and RPE6-C10

5.1 RPE01-D2

Clone RPE01-D2 was derived from the SSRPE cDNA library and found to show similarity to exon 4 and exon 5 of the human LOC145226 (later FLJ30273, now *RDH12*, accession number NM_152443) as shown in Figure 25. To establish the expression profile of RPE01-D2, Northern blot analysis was performed in bovine tissues using a 400-bp RPE1-D2 probe. A 2.1-kb was detected in bovine retina and RPE with abundant expression in retina, but not in bovine heart, liver, brain, kidney and lung (Figure 24). The retina specific expression of LOC145255 was further confirmed by RT-PCR analysis in adult human tissues (Figure 26). To isolate the full-length cDNA sequence, RT-PCR was performed. First-strand cDNA from human retina was used as a template using the primer pair RPE1-D2F/ RPE1-D2R (Figure 25). RT-PCR revealed three products, the larger is 950-bp in size. The RT-PCR products were ligated into pGEM® T-easy vector (Promega) and then transformed into ***E. Coli.*** Using blue/white colour selection. Positive clones were selected, and were sequenced as described. The sequencing revealed a 1112-bp transcript with an ORF of 951-bp. The presence of an in-frame stop codon 66-bp upstream of the putative start codon strongly suggests that the entire coding region has been isolated. The protein predicted from the ORF consists of 316 amino acid residues with a calculated molecular mass of 35.1 kDa. The exon/intron boundaries of human *RDH12* were determined by sequence alignment of cDNA clones to the genomic sequence. *RDH12* spans approximately 13.6 kb and consists of 7 exons and 6 introns (Table 11). The exons range in size from 68 to 210 bp. Intron sizes range from approximately 500 bp to 4.3 kb. All the 5`**-**donor and 3`**-**acceptor sites are consistent with the GT-AG consensus for pre-mRNA splicing recognition sequences. The gene maps to chromosome 14q23.3.

Table 11 Exon/intron Boundaries of the human *RDH12* gene

Exon	Exon size (bp)	Intron size (kB)	5` donor	3` acceptor
1	68	1762	CATCAG**gt**ttgtct	cgat**ag**GAAGT
2	119	507	GCCGAG**gt**aagt	accc**ag**GAGCCC
3	156	796	TGGCAG**gt**gagg	ctat**ag**AGGAAA
4	105	825	ACCTGG**gt**aagt	tcac**ag**GCCACT
5	210	2000	TCCAAG**gt**aagt	tccc**ag**GCACCG
6	190	4365	CTTCAG**gt**gtgt	ctcc**ag**TGACTG
7	103			

Note. The exon sequences are shown in uppercase letters, and the intron sequences are displayed in lowercase letters. The gt-ag consensus sequences in the splicing sites are shown in boldface type.

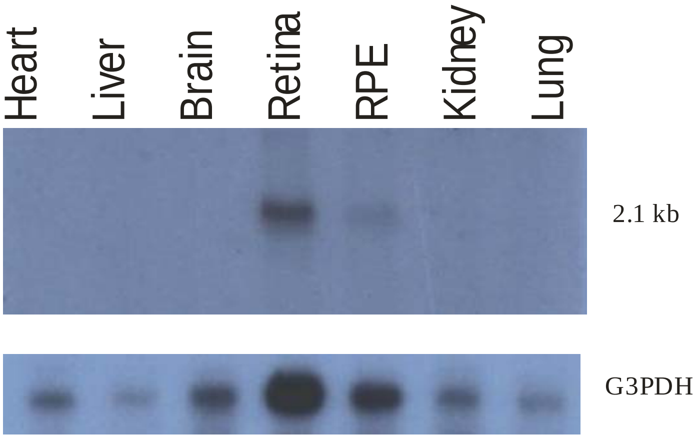


Figure 24 Northern blot hybridization of RPE01-D2 probe (400-bp) to bovine RNA samples. A signal about 2.1-kb was strongly detected in retina, weakly in RPE, but not in other tissues. G3PDH expression served as a control for cDNA integrity (Provided by Rahman).

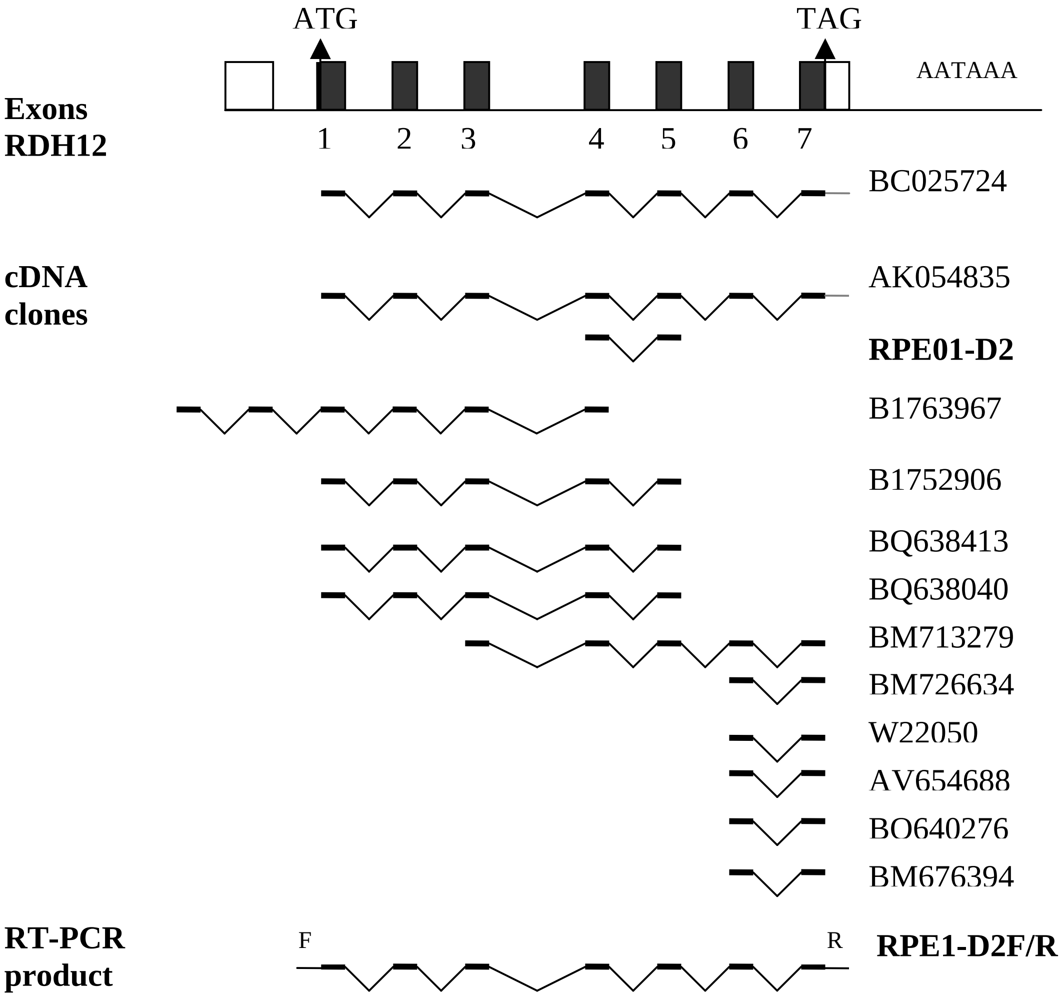


Figure 25 Schematic illustration of the exon-intron structure of the human RDH12 gene. The protein encoding exons are shown as black boxes and the untranslated regions as white boxes. Horizontal lines represent the sequence of cDNA clones and the retinal RT-PCR product used to assemble the full-length cDNA of the RDH12 gene.

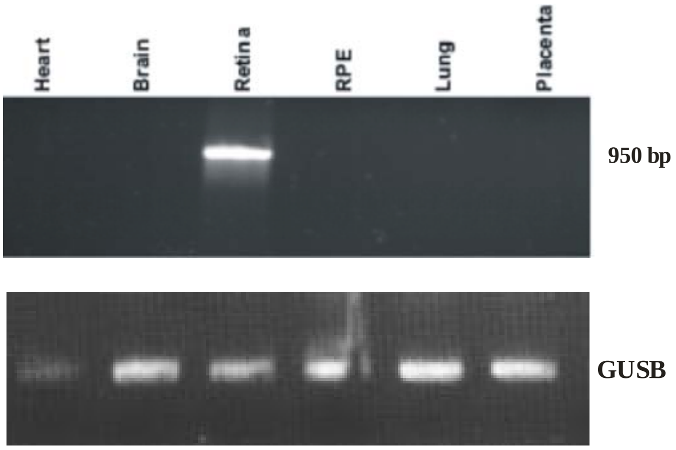


Figure 26 RT-PCR analysis of LOC145255 (RDH12) in human tissues with primer pair RPE1-D2F/R. The β-glucuronidase gene (GUSB) served as a positive control.

5.2 RPE06-C10

5.2.1 *Genomic structure and expression pattern*

32 clones from our SSRPE cDNA library were assembled into a 844-bp contig (ID 09), only 5 clones are shown in Figure 27. RPE06-C10 is one of those 32 clones, and therefore was used as a representative clone for this contig in expression experiments. The consensus sequence of this contig or cluster was found to show similarity to the 3` UTR of an intronless gene that is not annotated in the Human Genome Sequence draft (Figure 27). I will be referring to this single-exon gene as *6-C10*. Northern blot hybridization using RPE06-C10 clone as a probe and bovine RNA samples revealed that it is expressed highly in RPE and weakly in retina, but not in other tissues examined (Figure 28). The specificity of retina and RPE expression of *6-C10* was further confirmed by RT-PCR analysis in adult human tissues using primer pair RPE06-C10F3/RPE06-C10R3 (Figure 29). In addition, this gene was also found to be expressed in human lung.

This single exon gene is highly GC-rich (77%) and represents a highly conserved sequence between human, mouse and rat (Figure 27). GC rich sequences can be extremely difficult to RT-PCR amplify and to sequence. RPE06-C10 was amplified using 2X PCR$_x$ Enhancer solution using 3 pair of primers that were designed in an overlapping manner to cover the entire length of the single exon gene (Figure 27). Recently a newly identified Locus (LOC283211) that is located tail to tail with 6-C10 was included in the NCBI database (Figure 27). LOC283211 encodes also a single exon gene.

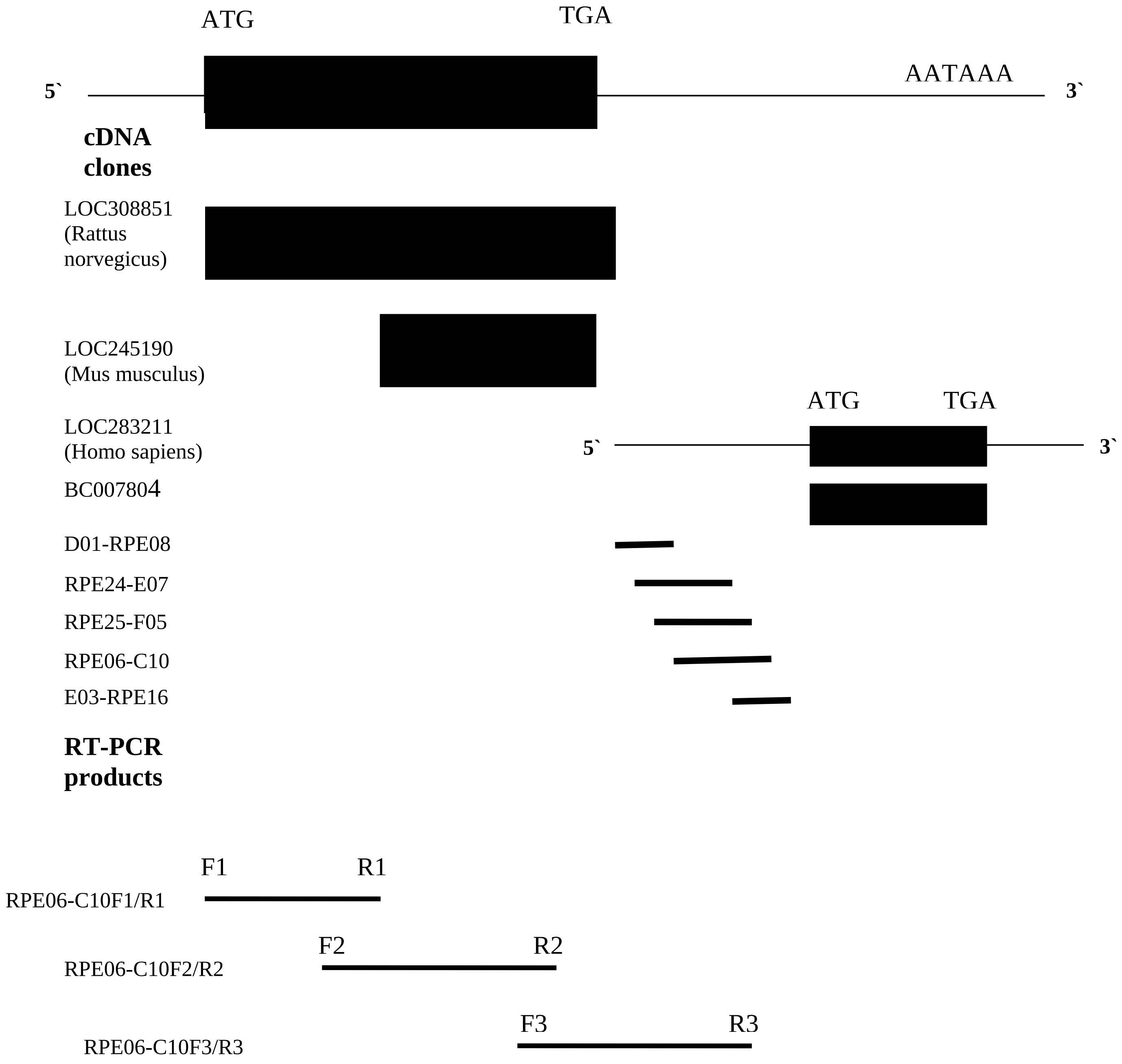


Figure 27 Schematic illustration of the structure of the human 6-C10 intronless gene. The protein encoding single-exon gene is shown as black box and the untranslated regions as black lines. Horizontal lines represent the sequence of cDNA clones and retinal RT-PCR products. Recently another new identified Locus (LOC283211) that has tail to tail with 6-C10 was included in the NCBI database. LOC283211 encoded a single-exon gene.

The genomic organization of *6-C10* was investigated by performing PCR amplification on genomic DNA template, using the 3 sets of oligonucleotide pairs encompassing the 5` and 3` ends of the *6-C10* mRNA transcript. The size of the resulting PCR products coincided with that of the *6-C10* cDNA used as template, demonstrating that cDNA corresponds fully to the genomic DNA suggesting that 6-C10 is an intronless gene. The gene maps to chromosome 11q13.3.

5.2.2 Analysis of the predicted cDNA sequence

A single long open reading frame of 1149 bases was identified, encoding a putative protein of 382 amino acids. The transcription start site was estimated to be 1953-bp upstream of the ATG start codon. The predicted molecular mass of *6-C10* is 40.4 kDa. Amino acid sequence comparisons to known proteins in the database indicate significant homology to the leucine rich repeat (LRR) domain. Proteins belonging to the LRR superfamily are thought to be involved in specific protein-protein and protein-matrix interactions (Buchanana and Gay, 1996).

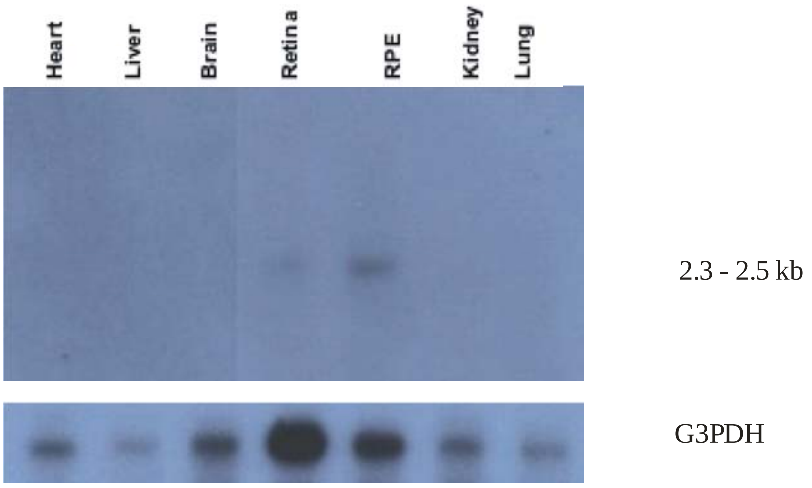


Figure 28 Northern blot hybridization of RPE06-C10 probe (600-bp) to bovine RNA samples. A signal about 2.3-2.5 kb was strongly detected in RPE, weakly in retina, but not in other tissues. G3PDH expression served as a control for cDNA integrity (Provided by Rahman).

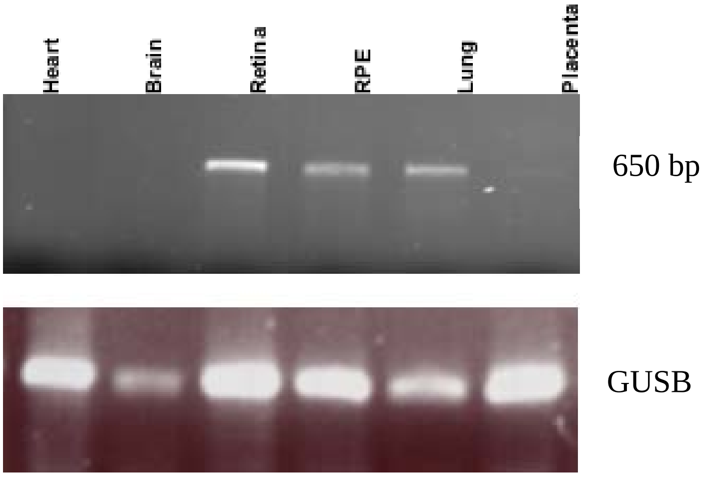


Figure 29 RT-PCR expression analysis of RPE06-C10 single-exon gene in human tissues with primer pair RPE06-C10F3/R3. The β-glucuronidase gene (GUSB) served as a positive control.

6. SNP genotyping for RDH12

To facilitate future association studies for the *RDH12*, SNP mapping was conducted using a combination of DHPLC and direct sequencing. A total of 25 fragments derived from the *RDH12* locus were PCR amplified and run on DHPLC or were directly sequenced. 12 SNPs were identified (Table 12). 5 SNPs showed allele frequency higher than 20% (frequency shown in bold and underlined in Table 12). 7 out of these 12 SNPs showed allele frequnecies less than 20% for the minor allele. 4 of the 7 lower allele frequency SNPs are in strong linkage disequilibium (shown with blue background in Table 12).

SNPs identified for the *RDH12* gene through the dbSNP were downloaded from GeneCards (http://bioinfo.weizmann.ac.il/cards-bin/carddisp?RDH12&search=DH12&suff=txt) and were compared to those identified. Table 13 summarizes the comparison between the dbSNPs and those identified in the present study. 8 out of the 10 dbSNPs were covered in the current investigation. Only 3 of those 8 were detected and only 2 of these 3 are of high frequency (≥20%). The identified SNP IVS6+539 corresponds to dbSNP (rs718212) by frequency (Table 13).

The identification of SNPs in the ***RDH12*** region was achieved by direct sequencing and dHPLC scanning. The main focus of SNP identification was within the coding region and exon/intron boundaries of the candidate gene, as well as 5 kb upstream and downstream of the 5’ and 3’ UTR, respectively. To increase the power to detect SNPs, we rescreened the ***RDH12*** gene in 48 individuals by direct sequencing. The region analyzed was a rescreening of the 23,400 bp encompassing 5 kb upstream of the coding region, thus it may contain regulatory elements of this gene. In addition, all the exons of ***RDH12*** were screened. 12 new SNPS were identified from direct sequencing of the 23.4-kb region.

Table 12 SNP card for *RDH12* gene

location	**PCR fragment**	**SNP ID***	**nucleotide change**	**Amino Acids change**	**frequency (allels)**	**fequency f**
5'UTR	1D2-5'UTR-F2/R2	-3924	-3924T>C		10/96	0.10
5'UTR	1D2-5'UTR-F5/R5	-3651	-3651T>C		10/96	0.10
5'UTR	1D2-5'UTR-F7/R7	-874	-874insAT		24/96	**0.25**
intron 2	1D2-ex2F/R	IVS2+54	IVS2+54G>A		7/96	0.07
intron 2	1D2-ex2F/R	IVS2+60	IVS2+60A>G		26/96	**0.27**
intron 2	1D2-ex2F/R	IVS2+151	IVS2+151G>A		1/96	0.02
intron 2	1D2-ex2F/R	IVS2+179	IVS2+179G>A		2/96	0.02
exon 5	1D2-ex5F/R	482	482G>A	Arg482Gln	10/96	0.10
intron 6	1D2-IVS6F1/R1	**IVS6+539**	IVS6+539T>C		28/96	**0.29**
intron 6	1D2-ex7F/R	**IVS6-260**	IVS6-260C>G		21/96	**0.22**
3'UTR	1D2-3'UTR-F2/R2	**2029**	+2029G>A		10/96	0.10
3'UTR	1D2-3'UTR-F7/R7	4819	4819C>T		31/96	**0.32**

*SNP ID is corresponding to the location of the identified SNP in the genomic sequence of RDH12 gene. For example, the SNP –3924 is located about 3924 bp upstream of the start codon of the gene, whereas the SNP 4819 is located 3868 bp downstream of the stop codon. The number 4819 is derived from adding the 3868 to 951 that represents the coding regions (cDNA) of the RDH12.

Table 13 dbSNPs for the *RDH12* gene

SNP	**Contig Accession**	**5' Flanking Sequence**	**3' Flanking Sequence**	**Validation**	**DNA Chg**	**AA Chg**	**Type**	**SNP ID**[d]	**Status in the present study**
rs2320030	NT_026437	CGTTCAAGGC	AAGACTTTAG	by-cluster[a]	A/G	--	**intron**	IVS4+430	not detected
rs718212	NT_026437	GGTGAAATCT	TTCCCATTC	by-frequency[b]	C/T	--	**intron**	IVS6+539	High frequency
rs2009590	NT_026437	TCACTCCTTG	TCTGTTGGTC	by-cluster	C/T	--	**locus-region**	-1660	not detected
rs910315	NT_026437	TAATCCACTC	TATTTTAAGC	by-cluster	A/G	--	**locus-region**	-437	not detected
rs756473	NT_026437	TGAAAGCCTC	GAAATGACCC	no-info[c]	A/G	--	intron	IVS1+1222	not detected
rs761512	NT_026437	acttctgctc	ccactacttt	no-info	C/G	--	intron	IVS6-260	High frequency
rs4899221	NT_026437	gtgcccagcc	TGATAGGCTT	no-info	C/T	--	intron	IVS6-865	not covered[d]
rs4899222	NT_026437	GACTTTTATA	TTAGAAAAAA	no-info	C/T	--	intron	IVS6-803	not covered[d]
rs742865	NT_026437	GTACAGGGCT	GGAGTGGATA	no-info	A/G	--	**locus-region**	2029	Low frequency
rs718213	NT_026437	GTATTTATTC	GTGAAGCACT	no-info	C/T	--	intron	IVS6+624	not detected

Lower case sequence letters indicate repetitive or low-complexity sequence.

[a] By cluster means that the SNP has been identified by comparison of human ESTs and genomic clones in the public databases.

[b] By frequency means that the SNP has been identified with Allele Frequency data.

[c] No information means that the SNP has been identified without submitting data about its frequency.

[d] For SNP ID definition see Table 13.

[e] not covered because they are just new to dbSNP database.

V DISCUSSION

The retinal pigment epithelium (RPE) is a single cell layer adjacent to the rod and cone photoreceptors that plays a key role in retinal physiology and the biochemistry of vision. Anatomically, RPE is located between the sensory retina and the choroid, this location subsequently exposes the RPE cells to a highly oxidative environment due to a high oxygen partial pressure from the underlying choriocapillaris. Physiologically, RPE cells phagocytose and digest photoreceptor outer segments. By virtue of its location, RPE forms the outer blood-retina barrier, which facilitates the selective transport of molecules between the outer neural retina and the choroidal blood stream. To meet the diverse and unique tasks in the continual support and renewal of the rod and cone photoreceptors, the RPE requires a large number of active genes some of which specific to these epithelial cells.

The aim of the present study was to identify and characterize genes expressed exclusively or abundantly in the RPE and to assess their contribution to AMD. Towards this end, a bovine cDNA library enriched for RPE-specific transcripts was constructed in-house using a suppression subtractive hybridization technique. This technique normalizes sequence abundance and achieves high enrichement for differentially expressed genes. The individual sequences of 2379 cDNAs derived from the SSRPE cDNA library were analyzed and querried against sequences deposited in GenBank and dbEST databases. This process identified 341 distinct genes represented by 1748 ESTs. Some of these have been confirmed to be involved in retinal dytrophies such as ABCA4, EFEMP1, RDS, TIMP, VMD2 and RPE65.

1. Candidate gene approach and SNP analysis

During the last few years, the identification and characterization of disease genes has become the main approach to understand the role of specific genes in the pathogenesis of inherited disorders. Several strategies have been developed to identify these disease genes including the positional cloning and the positional candidate gene strategy. This approach allowed the exclusion of several genes as a candidate locus of AMD, VMD2 (Krämer et al., 2000); EFEMP1 (Guymer et al., 2002); TIMP3 (De La Paz et al., 1997) and lead to the implication of two genetic factors: the apoE gene (involved in the transport of lipids) and the ABCR gene (involved in Stargardt macular dystrophy).

Candidate gene approaches can utilize either linkage analysis of highly polymorphic, short tandem repeat marker loci in the candidate region, or association of coding or noncoding polymorphisms within the candidate genes. The late onset of the disease and the fact that AMD is a polygenic and multifactorial complex disease are the main limiting factors for linkage analysis studies (Dryja, 1997). Genes expressed in a certain tissue can be assumed to be important for the function of that tissue. Once isolated, genes that are highly expressed or tissue specific can become candidate genes for disorders that affect the tissue used for gene isolation. For example, genes expressed in the retina can be candidate genes for retinal disorders.

In order to handle and evaluate the massive amount of data generated by the SSRPE cDNA library, a highly efficient and user-friendly database management system was developed using Microsoft Access.

2. RDBMS construction

The development of Access applications requires several steps. The most important part of the application process is the development of an efficient and maintainable database utilizing logical data modeling and normalization. Normalization simply means reducing data to their simplest elements, and only storing each element once. It is important to note that most applications fail due to mistakes in analysis and design, not because of mistakes' construction. During design phase of the application development process, entities will be used to derive the groupings of data which will be physically stored, e.g. tables, record tpyes, and files. For each entity there has to be one attribute or a group of attributes which will uniquely identify each occurrence of the entity. This attribute (or group of attributes) is the entity's primary key. For example, consider the entity CLONE. The system needs to be able to uniquely identify each occurrence of this clone entity. Clone Name can be the primary key as for this example, it is not possible for two or more clones to have the same Clone name. A clone name will always identify a single clone and eventually a single occurrence of the entity CLONE. In some cases, no single attribute is sufficient on its own to form the primary key of an entity. In such cases, it has to be considered whether there are two or more attributes which together can form the primary key. For example, assume the entity CLONE – GENE which contains details of a clone similarity. Each of the attributes on its own is insufficient to act as the primary key of the entity. However, attributes Clone Name and Gene ID together will identify an individual clone similarity; therefore these two attributes become the primary key of the CLONE – GENE entity. In such cases, the primary key is called a compound key consisting of two or more attributes, each of which either individually or in combination is the primary key of other entities. Both primary and compound keys are shown as bold character in the entities and attributes for DBMS.

MS Access is a relational database management system (RDBMS), which stores and retrives information according to relationships defined by the analyst. It can help organizing data to make it easier to enter, edit, and retrieve the information. With Access, *tables* can be created to contain information, *queries* to retrieve information from tables, and *forms* and *reports* to make the information available to users in various ways. VBA (Visual Basic for Applications) brings the power of Visual Basic to Access application. Access is extremely powerful by itself but occasions arise where Access does not have the necessary tools to create some of the features that are required by the application. One of the main purposes of a database is to store information. Tables, relationships, and other techniques were designed and implemented to ensure that the data are maintained in the way that such a database requires. However, these data are virtually useless unless we can retrieve the information and manipulate it in whatever way needed. A *query* is a question that can be asked in a language called SQL. From a database of genes, we might want to know, for example, how many of them are RPE-specific and how many are retina-specific. Queries allow us to formulate these questions for the constructed database and view the results. It is important to understand that queries do not return tables. Rather, they return ***recordsets***. Although they may seem similar, queries and recordsets are quite distinct. Queries are the *questions* we ask the database in SQL language; the recordsets are the *answers* that we get back from the database. An effectively designed database will be easier to expand as the requirements of the database's information grow and change. For example, at the beginning of this project we searched our bovine RPE ESTs sequences against Blast nt, htgs, and dbEST. During the course of this project the draft of the human genome was launched to public access (Feburary 2001). Consequently, our database was expanded easily and effiencely to accommodate the new developments.

3. SSRPE cDNA library analysis

To understand the molecular structure and function of the RPE, the relevant subset of differentially expressed genes needs to be identified, and studied in detail. The identification of the genes expressed in the RPE constitutes a necessary step towards the understanding of its physiology. A powerful approach to the analysis of expression in a particular tissue or cell type involves single-pass partial sequencing of clones from a cDNA library together with the comparative analysis of the resuling ESTs with entries in public databases. Because cDNA libraries are typically generated from tissues or developmental stages which are randomly selected for sequencing, EST representation provides a dynamic view of genome content and expression. EST sequencing has been widely used as an efficient approach for gene discovery. It should be noted, however, that libraries generated using standard PCR methods to include a number of artefacts because of over-cycling of the PCR reaction. Also, such libraries may result in an enrichment of small cDNA clones that do not possess the entire open reading frame (ORF) of a gene.

To construct a high-quality RPE cDNA library, RPE cells were required in large quantities as well as excellent quality, free of any contaminating cells from the adjacent retina or the choriod. These requirements can not be met with human RPE. In the present study, a PCR-based SSH method (Diatchenko et al., 1996, 1999) for cDNA subtraction was performed to isolate and identify mRNAs from bovine RPE. This approach normalizes sequence abundance and achieves high enrichment for differentially expressed cDNAs by single rounds of subtraction against a mixture of cDNAs from several non-ocular tissues (bovine heart and liver). Subsequently the expression pattern of individual clones of the RPE cDNA library can be determined. RPE clones expressed in RPE and/or retina but not in other tissues are being selected for further analysis. Genes expressed in a certain tissue can be assumed to be important for the function of that tissue. Clones found to be more highly expressed in RPE and/or retina than in other tissues may also choose for further analysis. The genes resulting in retinal degeneration which, in some case is a component of a more complex phenotype are typically ubiquitously expressed. For example, mutations in TIM3, a widely expressed human gene, lead to Sorsby's fundus dystrophy that is characterized by accumulation of lipid deposits in Bruch's membrane. (Weber et al., 1994). De La Paz have excluded the TIMP3 gene as a cause for AMD (De La Paz et al., 1997).

In a first phase, 1002 subtracted bovine RPE cDNA clones were sequenced (in-house) and blasted to available sequences in the non-redundant GenBank/EMBL and dbEST databases. The 1002 bovine RPE-ESTs represent 168 known human genes and approximately 107 predicted or as yet unknown genes. Overall, a high degree of conservation between bovine and human sequences was noted and was in the range of 80 to 92% for coding and 70 to 85% for non-coding (mostly 5` or 3` UTRs) regions. A detailed comparison of randomly chosen RPE-EST clones 1-200 with RPE clones 800-1000 demonstrated that the latter sample group still contained a high number of novel sequences suggesting that the subtracted RPE cDNA library was not exploited exhaustively, and further bears the potential for the identification of additional novel RPE transcripts. *In silico* expression profiling of the 168 known genes was performed to evaluate the SSRPE cDNA library for the degree of enrichment. Data were obtained for 153 genes, revealing that about 17% (26/153) of the reported genes were RPE and/or retina-specific and around 5% (7/153) were specifically expressed in retina and brain. Extrapolating these data has led us to predict that the approximately 107 unknown genes identified in the first phase of this study could include 15-20 novel RPE/Retina-specific genes. Northern blot analysis was performed for the 107 predicted or as yet unknown genes using bovine RNA samples (Faisal Rahman, personal communication). The results of the expression have demonstrated that 24 out of the 107 clones (23%) analyzed show exclusive or abundant expression in the ocular tissues, in particular the RPE and retina. Further analysis of these 24 clones revealed that 8 were either redundant clones or show similarity to known genes. The remaining 16 RPE and/or retina-specific clones representing predicted or as yet unknown genes represent excellent candidates for association studies in AMD/control cohorts. Of the 16 clones, 2 were further characterized in the present study.

In a second phase, an additional 1377 subtracted bovine cDNA clones were sequenced (Lynkeus) and blasted against a recent release of the human genome draft sequence. This raised the total number of the bovine RPE cDNA clones to 2379. CAP3 (Huang and Madan, 1999) was used to assemble the high quality sequences of all the 2379 ESTs into clusters or singletons. 1.2% of the 2379 RPE-ESTs contains vector sequences and was excluded from further analysis. 5% of the RPE-ESTs showed homology to multipe chromosomes and were not included in further assembly process. The rest of the ESTs (2245) were assembled into 175 contigs and 509 singletons, which revealed approximately 684 unique genes in the dataset. Of these, 341 were attributed to known or predicted genes and 343 transcripts did not show human orthologues. The frequency of EST distribution after CAP3 analysis and BLAST searches showed that most clusters contained 1-3 ESTs and the largest cluster contained 151 ESTs (Table 9).

Among the top most abundant ESTs (Table 9) are the genes that have been previously shown to be highly expressed in RPE, e.g. the VMD2 gene causing Best vitelliform macular dystrophy which is an autosomal dominant disorder (Marquardt et al., 1998; Petrukhin et al., 1998) is represented by 151 ESTs. In addition, we assessed the validity of the subtracted library by examining ESTs from genes known to be specifically expressed in other ocular tissues than the RPE. For example, the genes for RHO, RDS, and RCV1 are known to be expressed specifically in the retina playing a role in phototransduction, but were found in our library (Table 9). This can be considered as retinal contamination of the SSRPE cDNA library; although it may be expected as retina and RPE cannot be separated completely. It has been estimated that RPE RNA contains 1% retina RNA and vice versa (den Hollander et al., 1999). Retinal degeneration might happen because abnormal proteins lead to death of the outer segments (OS), but it might also occur because abnormal OS cause damage to the RPE after they are phagocytized and digested. The subsequent metabolic damage to the RPE may lead to changes that cause further photoreceptors death. In another example, Stargardt's disease is characterized histopathologically by an accumulation of lipofuscin-like material throughout the RPE, but the recently isolated ABCA4 gene codes for a lipid transporter in the rod outer segments. It seems that the photoreceptors abnormality leads to the accumulation of abnormal material in the RPE (Marmor, 1998).

The main reason that about 343 RPE transcripts did not show human orthologues might be that these clones might include the 3’ UTR of the gene and not the coding regions (Sharma et al., 2002). It is the coding regions that can be conserved between bovine and human and not the 3’ UTR. This lacks of human ortholgous can be resolved using two approaches. First, continuous blasting might reveal their human orthologoues. For example, the clone RPE24-D11 that was found to be RPE specific and therefore selected as an AMD candidate gene, used to show no significant similarity from May 2002 until March 2003. Recent blast searches (July 2003) revealed that clone RPE24-D11 shows similarity to LOC119587 (accession number XM_058409). Secondly, more sequencing from the cDNA library together with reclustering of those 343 ESTs might reveal their human orthologoues. However, this does not mean that all of them are going to be novel, because some of them might show similarity (after reclustering) to already reported human genes in the database. The quality of the SSRPE cDNA library is reliable and its construction is the basis for further screening differentially expressed genes of RPE. Our results show that suppression subtractive hybridization (SSH) is a suitable method for identifying rare transcripts.

4. Blast searches and gene predictions

In the present study, genes were identified on the basis of the following criteria: identities or similarities to known proteins; identities to spliced ESTs; and patterns of consistent coding-exon prediction. First, protein matches identified genes that were identical or similar to known genes in the public databases. Secondly, for EST matches, only those that showed evidence of splicing were used, those that were non-contiguous with genomic sequence, showed consensus splice sites, and represented essentially perfect matches (>95% identity) to the genomic sequence. Finally, the criterion of consistent exon prediction required that two of the coding-exon prediction programs (Grail, Genscan) agreed on the location of the exon.

5. From genes to functions

Analysis of the SSRPE cDNA library has provided a first look at the transcriptome of this tissue. The next steps require isolating complete cDNAs for each gene. With complete coding sequences, protein sequences can be deduced and examined for motifs, domains, and biochemical characteristics that may suggest function. The most challenging problem will then be to determine the functions of these genes. While it is tempting to focus on genes whose protein characteristics suggest a hypothesis for relevance to some aspect of AMD or other retina dystrophies, the more than 343 transcripts identified during the course of this study that have no functional association are too large a dataset to ignore. For these and other genes, detailed expression analysis may be informative. Completion of the expression analysis experiments is in progress using real time RT-PCR and will provide additional candidate genes for further analyses. Demonstration that a gene shows increased expression in RPE and/or retina by Northern blot or RT-PCR analysis, followed by RNA tissue *in situ hybridization* to define specific cell types and developmental stages of expression, may help in selecting genes of greater or lesser interest.

During the course of the present project, we have identified 341 known genes. A variety of tools, including BLASTN, BLASTX, LocusLink, SwissProt, Source, Pfam and published literature were used for the functional annotation of the known and predicted genes. This revealed 18 different functional groups (Figure 22). Of the 341 known genes, no function could be assigned to 24.4%. In some cases, the lack of protein or functional domain data may be due to the lack of complete coding sequence information. Genes involved in metabolic pathways are represented and corresponded to 14.5% of the total number of annotated genes, although it may be expected as the RPE is a highly active metabolic tissue. 12.5% of the known genes involved in various types of transport. For instance, SLC6A13 (solute carrier family 6 member 13) is involved in transport. SLC6A13 is highly expressed in RPE/choroid, and was identified in the current library. We also found that 10.3% of the known genes are involved in cell-cell signalling, e.g. secreted frizzled-related protein 5 (SFRP5), which might act by modulating Wnt signalling transduction and is highly expressed in RPE, and is also expressed in pancreas (Jinghua et al., 1999). Recently, it was found that disruptions of Wnt network signalling are involved in retinal neurodegeneration, and that targeting of SFRPs to key areas of the neuroretina may mediate mechamisms promoting cell death (Jones et al., 2000).

6% of the known genes fall into the cell defence group as RPE is subjected to high oxidative stress. This is due to the high partial pressure of oxygen, the high metabolic rate of the RPE, accummulation of lipofuscin and chronic light pressure. Subsequently, an increase level of reactive oxygen species (ROS) would distrub its clearance and result in oxidative damage to macromolecules (Ames et al., 1993; Harman, 1998). To cope with these toxic oxygen intermediates, the RPE is protected by an effective defence mechanism against oxidative damage. It is particularly rich in anti-oxidants such as glutathione-S-transferases, and glutathione (Beatty et al., 2001). We identified in our library several antioxidants belonging to multiple isoforms of the gene glutathione-S-transferses e.g. GSTM1, GSTM5.

Within the eye, several transcription factors are known to be critical to proper ocular development. Transcription factors have been reported to regulate gene expression by binding directly to promoter sites within DNA or to other transcription factors (Jobling et al., 2002). 4% of the reported genes are involved in transcriptional factors, including OTX2 (orthodenticle homolog 2 Drosophila), and NRL (neural retina leucine zipper). Among many transcription factors that affect eye morphogenesis, OTX2 is of interest, because it is expressed not only at an early stage of eye morphogenesis, but also in RPE of postnatal and the adult mouse eye (Baas et al., 2000). NRL was initially identified as a retina specific and was subsequently shown to be expressed throughout the retina cells (Swaroop et al., 1992).

3% of the identified known genes in this study are involved in vitmain A (retinol) metabolism and transport, including RPE65, RLBP1, RBP1, LRAT, BCDO1, RBP4, and RDH5. RPE has diverse features including the uptake, processing, transport and release of vitamin A (retinol) and some of its visual cycle intermediates (retinoids). Retinol uptake occurs at both the basolateral and apical surfaces. Delivery of retinol across the basolateral membrane is mediated by a retinol binding protein (RBP) that is secreted by the liver as a complex with retinol (vitamin A). Within the cell, retinol and its derivatives are solubilized by intracellular retinoid binding proteins that are selective for retinol (cellular retinol binding protein, CRBP) and 11-cis retinoids (cellular retinal binding protein, CRALBP).

Protein degradation represented 3% of the reported genes; of this 1% are lysosomal enzymes. Lysosomal enzymes are responsible for the degradation of a variety of proteins, lipids, nucleic acids and complex carbohydrates. Lysosomal enzymes are much more active in RPE than in many other tissues of the body and are responsible for the degradation of ingested photoreceptor outer segments (Verdugo and Ray, 1997). Alteration of any of these enzyme activities with aging might be involved in the pathogenesis of AMD.

2% of the reported genes are involved in phototransduction. Phototransduction takes place in the outer segments of the rod and cone photoreceptor cells. The small percentage of photoreceptor-specific genes in the RPE could be considered as contamination of the RPE with retina.

The known genes identified in this RPE library reflect our knowledge of RPE physiology and may facilitate better understanding of human macular degeneration. It is expected that the unknown transcripts will also shed further light on the function and physiology of this important ocular tissue.

6. Comparison to similar studies

EST analysis of a cDNA library is a powerful tool for probing the transcriptome of tissues and cell types. There are two approaches for cDNA libraries construction either un-amplified and unnormalized or amplified and subtracted. On one hand, the unamplified and unnormalized cDNA library is constructed to have as close as possible a representation of normal transcript abundance, maximize clone length and to allow the discovery of difficult clones that might disappear during library manipulation (Wistow et al., 2002b). On the other hand, cDNA library is amplified and subtracted to equalize the abundance of cDNA within the target populations and exclude the common sequences between the target and driver population (Diatchenko et al., 1996; 1999). At the initiation of the present project (May 2000), only two studies reported the identification of RPE-specific genes using subtracted cDNA libraries from human RPE-cell line (Gieser and Swaroop, 1992) and human RPE and choroid (den Hollander et al., 1999). Since then and during the course of the current project, other similar studies have been reported (Sharon et al., 2002; Buraczynska et al., 2002; Wistow et al., 2002a; Sharma et al., 2002). Of the approximately estimated 10000 transcripts in the mammalian RPE (Swaroop and Zack, 2002), only 1-2% are estimated to be unique or differentially expressed transcripts (Zhang et al., 1997). The transcripts that specifically or differentially expressed in the human RPE may be estimated to be about 100-200.

Sharma et al. reported sequencing and analysing of 1000 clones from a subtracted bovine RPE cDNA library (Sharma et al., 2002). Northern blot analysis was performed for 100 non redundant clones. 45 of the clones gave positive hybridization, and 2 of these 45 were present in duplicate and excluded from further analysis. 19 of the remaining 43 clones showed homology to known mammalian genes, whereas 24 of them showed no matches to any known gene. Comparison of the results of Northern blot analysis of 91 non redundant clones from the Sharma library to the 107 predicted genes from our SSRPE library revealed the tissue distribution of 50% of the cDNA clones (Table 14). Expression of half of the subtracted cDNAs was not detectable under the conditions of Northern hybrization used in both studies, possibly suggesting a low level of expression of their respective transcripts (Table 14). Northern blot analysis of poly-A+ RNA may reveal the transcipts and tissue distribution of these cDNAs (Sharma et al., 2002). The expression analysis demonstrated that the subtractive hybridization method used in both studies enabled the enrichment of genes expressed in the bovine RPE.

Wistow et al. reported that unamplified, un-normalized RPE/choroid (Wistow et al., 2002a) and retina (Wistow et al., 2002b) library reveal an abundantly expressed genes. These libraries also revealed cDNAs that may be selectively lost through library manipulations such as amplification because of poor growth characteristics in bacterial hosts (Wistow et al., 2002b). The normalized version of the same libraries loses abundance information, but has the potential for identification of the rarer transcripts.

Table 14 Summary of the expression of 107 SSRPE cNDAs and 91 subtracted bovine cDNAs from the Sharma RPE library

Tissue specificity	**SSRPE**[a]	**Sharma**[b]
RPE-specific	7	11
Retina-specific	3	13
RPE/retina specific	7	11
Tissue-restricted	7	7
Ubiquitous	29	3
No possible evaluation	54[c]	46[c]
Total	**107**	**91**

[a]Data provided by Faisal Rahman (Personal commincation, 2002).

[b]Sharma et al., 2002

[c]Expression of about half of the subtracted RPE cDNAs was not detectable under the conditions of Northern blot hybridization possibly suggesting a low level of expression.

Wistow et al. (2002a) generated 10,000 ESTs from the human RPE/choroid for the NEIBank project. The NEIBank database was created by the National Eye Institute (Wistow, 2002c; http://neibank.nei.nih.gov/index.shtml). Comprehensive analysis of the NEIBank RPE/choroid unique gene clusters revealed that 3820 attributed to known genes in the public databases and the rest of 2480 have no matches. Comprehensive comparison of the 3820 unique RPE clusters with our datasets of 341 RPE known genes shows that only 146 are overlapping between the two libraries (Table 10 & Figure 23). 42.8% of our identified RPE genes are overlapping with only 3.8% of NEIBank RPE/choroid known genes. This result is not surprising, as the NEIBank genes are from combined RPE and choroid, many transcripts might have been derived from ocular vasculature (Wistow et al., 2002a). In addition, the RPE/choroid library was an unamplified and unnormalized library, whereas our cDNA library is subtracted, suggesting that many of these 3820 uniqiue RPE could be house-keeping genes.

Further analysis of the 146 overlapping genes shows that 11 of them are involved in retinal dystrophy (shown as bold and underlined in Table 10). For example, the EGF-containing fibrillin-like extracellular matrix protein-1 (EFEMP1) gene is associated with Malattia Leventinese (ML) and Doyne honeycomb retinal dystrophy (DHRD) (Stone et al., 1999). ML and DHRD refer to two autosomal dominant diseases characterized by yellow-white deposits known as drusen that accumulate beneath the

RPE. The EFEMP1 gene is expressed in the retina, RPE and choroid and shows homology to a family of extracellular matrix glycoproteins known as fibulins. The Arg345Trp disease-associated allele of the EFEMP1 gene has been excluded from been associated with AMD etiology (Guymer et al., 2002). However, this does not exclude the involvement of other alleles of the EFEMP1 gene in the AMD phenotype.

7. Cloning and characterization of RPE01-D2 and RPE6-C10

As was mentioned previously, 16 RPE and/or retina-specific clones representing predicted or as yet unknown genes, were selected as AMD candidate genes. To further evaluate these clones extensive homology searches for protein motifs and assignment to established protein families have been performed. This resulted in cloning and characterization of two clones, RPE01-D2 and RPE06-C10.

7.1 RPE01-D2

The clone RPE01-D2 was identified from our subtracted RPE cDNA library and found to show similarity to exon 4 and exon 5 of the human LOC145226 (later FLJ30273, now *RDH12*, accession number NM_152443). RPE01-D2 was found using Northern blot hybridization to be expressed in bovine retina and RPE with abundant expression in retina, but not in other bovine tissues examined. The retina specificity of LOC145255 was further confirmed by RT-PCR analysis in adult human tissues.

Amino acid sequence comparisons of *RDH12* to known proteins in the database indicate significant homology to the short-chain dehydrogenases/reductases (SDR). The SDR is a very large family of enzymes, most of which are known to be NAD- or NADP-dependent oxidoreductases (Joernvall et al., 1995). Most members of this family are proteins of about 250 to 300 amino acid residues. The amino acid sequence of *RDH12* contains two motifs highly conserved among the SDRs superfamily, the cofactor-binding site (GXXXGXG) and catalytic residues (YXXXK). The first binds the coenzyme, often NAD, and the second binds the substrate. The latter domain determines the substrate specificity and contains amino acids involved in catalysis. Recently, strong signals of the *RDH12* transcript in photoreceptor inner segments of monkey and mouse retina has been detected using *in situ hybridiation* (Haeseleer et al., 2002). Although the function of *RDH12* is at this point unclear, it was suggested that it could be involved in the formation of 11-cis retinal from 11-cis retinol during regeneration of the cone visual pigments (Haeseleer et al., 2002).

The *RDH12* gene maps to chromosome 14q23.3 and is located 30 kb from the *RDH11* gene in the 3'-RDH11-5' 5'-RDH12-3' orientation. *RDH12* and *RDH11* genes are located at the locus for an autosomal recessive retinal distrophy, Leber's congenital amaurosis 3 (LCA3). This autosomal recessive disorder is usually recognized at birth or during the first months of life in an infant with total blindness or impaired vision. Recently, mutations in three genes have been associated with LCA phenotype. These include a retinal-specific guanylate cyclase (Perrault et al., 1996); RPE65 (Marlhens et al., 1997); and CRX (Freund et al., 1998). Additionally, several recent studies have suggested that LCA is genetically heterogeneous and still there are genes remain to be identified for this disease (Stockton et al., 1998 ; Hameed et al., 2000). Mutations in CRALBP have been linked to cases of autosomal recessive retinitis pigmentosa (Maw et al., 1997); and mutations in 11-cis-retinol dehydrogenase have been reported in patients with fundus albipunctatus (Yamamoto et al., 1999). *RDH12* could be considered as a candidate gene for AMD and other retinal dystrophies due to the fact that many retinal degeneration have been associated with mutations in genes involves in the visual cycles and retinoid metabolism.

7.2 RPE06-C10

6-C10 is an intronless gene and thus far has not been annotated in the Human Genome Sequence draft (Build 33). Northern blot analysis using RPE06-C10 clone as a probe and bovine RNA samples revealed that it is highly expressed in RPE and weakly in retina, but not in other tissues examined. The retina and RPE specificity was further confirmed by RT-PCR analysis in adult human tissues. In addition, this gene was also found to be expressed in human lung.

This intronless gene is GC-rich and represents a highly conserved sequence between human, mouse and rat. GC rich sequences can be difficult to amplify by PCR. 6-C10 was amplified using 2X PCR$_x$ Enhancer solution and 3 pairs of overlapping primers that cover the entire length of the single exon gene. Because of its highly GC-rich, 6-

C10 illustrates some of the features that may be common to other genes that are missing from the current builds of the human genome, making it difficult for reverse transcription and for growth in bacterial host cells (Wistow et al., 2002a).

The genomic organization of the *6-C10* was investigated by performing PCR amplification on genomic DNA template, using the 3 sets of oligonucleotide pairs encompassing the 5` and 3` ends of the *6-C10* mRNA transcript. The size of the resulting PCR products coincided with that of the *6-C10* cDNA used as template, demonstrating that cDNA was corresponding in full to the genomic DNA suggesting that 6-C10 is an intronless gene. A single long open reading frame of 1149 bases was identified, encoding a putative protein of 383 amino acids. It is noteworthy that the coding exons of intronless genes tend to be very large generally encompassing more than one kilobase (kb) in length, in contrast to typical coding exons that average 100-150 bp. From the large open reading frame of *6-C10*, a protein of 383 amino acids was deduced, with a predicted molecular mass of 40.4 kDa. Amino acid sequence comparisons to known proteins in the database indicate significant homology to 3 leucine rich repeats (LRR) and one LRR C-terminal domain (LRRCT). LRR domains are short sequence motifs present in a number of proteins with diverse functions and cellular locations. LRRs are often flanked by cysteine rich domains. This domain is often found at the C-terminus of tandem leucine rich repeats. LRR proteins have been reported in various species with diverse functions and proteins belonging to the LRR superfamily are thought to be involved in specific protein-protein and protein-matrix interactions (Buchanana and Gay, 1996). Recently, a novel β-Amyloid-induced (Aβ) rat and human gene, designated as Lib and LRRC15 respectively, suggests to be a member of the LRR superfamily which is involved in cell-cell and/or cell-extracellular matrix interactions including adhesion or target recognition in neuroinflammation states (Satoh et al., 2002). The availability of a mouse orthologue of the *6-C10* gene makes its functional analysis feasible through the generation of transgenic mice.

8. Generation of a SNP map of RPE01-D2 (*RDH12*)

Genotyping involves the analysis of genomic DNA to identify naturally occurring differences between individuals. These DNA differences, referred to as polymorphisms, can be used as “molecular signposts” to locate genes involved in the development of complex diseases such as diabetes, asthma, or AMD. Disease genes are identified by defining genetic linkage between polymorphisms and the distribution of disease within affected populations. A statistically significant link between known polymorphisms and the disease indicates that a disease-contributing gene is located within the same chromosomal region of the polymorphism. Thus, genotyping facilitates gene identification for drug target discovery and identification of diagnostic markers.

In addition to SNP genotyping, a novel approach to SNP haplotyping is the process of determining the presence of linked SNPs. Haplotypes result from the simultaneous presence of two or more DNA variations or SNPs that are genetically linked. Haplotyping is gaining increased attention because multiple linked SNPs have the potential to provide significantly more power for genetic analysis than individual SNPs. Thus, haplotype tag SNPs (htSNPs) may be used more efficiently for disease association studies to compare the genomes of diseased populations to matched control populations, and are expected to be more useful as diagnostic or predictive markers than individual SNPs.

In the present study, the identification of SNPs in the ***RDH12*** locus was achieved by direct sequencing and dHPLC scanning. The main focus of SNP identification was within the coding region and exon/intron boundaries of the gene, as well as 5 kb upstream and downstream of the 5′ and 3′ UTRs, respectively. 12 SNPS were identified from direct sequencing of the 23.4-kb region, of which only 5 are of high frequency and were identified in intronic regions of ***RDH12***. However, before any conclusions can be made, comparison of allele frequencies between patients with AMD phenotype and healthy controls is required.

VI CONCLUSIONS AND FUTURE DIRECTIONS

AMD is the leading cause of severe central visual impairment among the elderly and is associated both with environmental and genetic factors. Identification of the conferring susceptibility to AMD will facilitate an early diagnosis of the disease and may lead to the development of new strategies for prevention and therapy.

Although the generation of ESTs constitutes an efficient strategy to identify genes, there are some limitations to their approach. Firstly, it is difficult to isolate mRNA from some tissues and cell types. This results in a paucity of data on certain genes that may only be found in these tissues or cell types. However, this problem has been resolved in the present study by using bovine instead of human RPE. Secondly, important gene regulatory sequences may be found within an intron. Because ESTs are small segments of cDNA, generated from an mRNA in which the introns have been removed, much valuable information may be lost by focusing only on cDNA sequencing. Despite these limitations, the present study demonstrated that ESTs are invaluable in characterizing the human genome, as well as the genomes of other organisms. They have enabled the mapping of many genes to chromosomal sites and have also assisted in the discovery of many new genes. In this EST project, we used a simple and more direct approach to assess the RPE transciptome. Both *in silico* analysis and experimental data provided evidence that the SSRPE cDNA library is indeed enriched, indicating the high efficiency of the cDNA subtraction strategy.

The combination of bioinformatics analysis and laboratory investigation is an effective approach for the identification of candidate genes that are specifically or abundantly expressed in tissue-restricted manner (e.g. RPE/retina). Using genetic association studies, the contributions of individual genes to complex diseases that have a polygenetic basis like AMD, can be identified. By comparing frequency of SNPs in patients and controls, genes that contribute to disease can be determined. The future goal of this project is to develop comprehensive SNP maps of RPE-associated genes to enable extensive AMD case/control association studies.

VII REFERENCES

Algvere, P. V., P. Gouras and K. Dafgård (1999). Long-term outcome of RPE allografts in nonimmunosuppressed patients with AMD. *European Journal Of Ophthalmology* **9**(3): 217-230.

Allikmets, R., N. F. Shroyer, N. Singh, J. M. Seddon, R. A. Lewis, P. S. Bernstein, A. Peiffer, N. A. Zabriskie, Y. Li, A. Hutchinson, M. Dean, J. R. Lupski and M. Leppert (1997). Mutation of the Stargardt disease gene (ABCR) in age-related macular degeneration. *Science* **277**: 1805-1807.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.

Altschul, S. F., T. L. Madden and A. A. Schäffer (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389-3402.

Ames, B. N., M. K. Shigenaga and T. M. Hagen (1993). Oxidants, antioxidants, and the degenerative diseases of aging. *Proc Natl Acad Sci USA* **90**: 7915-7922.

AREDS Group (2000). Risk factors associated with age-related macular degeneration. A case-control study in the age-related eye disease study: age-related eye disease study report number 3. Age-Related Eye Disease Study Research Group. *Ophthalmology* **107**(12): 2224-32.

Baas, D., K. M. Bumsted, J. A. Martinez, F. M. Vaccarino, K. C. Wikler and C. J. Barnstable (2000). The subcellular localization of Otx2 is cell-type specific and developmentally regulated in the mouse retina. *Brain Res. Mol. Brain Res* **78**: pp. 26–37.

Beatty, S., I. J. Murray, D. B. Henson, D. Carden, H. Koh and M. E. Boulton (2001). Macular pigment and risk for age-related macular degeneration in subjects from a Northern European population. *Invest. Ophthalmol. Vis. Sci.* **42**: pp. 439–446.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp and D. L. Wheeler (2000). GenBank. *Nucleic Acids Res.* **28**: 15-18.

Bird, A. C., N. M. Bressler, S. B. Bressler, I. H. Chisholm, G. Coscas, M. D. Davis, P. T. De Jong, C. C. Klaver, B. E. Klein, R. Klein and e. al. (1995). An international classifcation and grading system for age-related maculopathy and age-related macular degeneration. The International ARM Epideniological Study Group. *Surv. Ophthalmol.* **39**: 367-374.

Bok, D. (1993). The retinal pigment epithelium: a versatile partner in vision. *J. Cell Sci Suppl* **17**: 189-95.

Bracey, L. T. and K. Paigen (1987). Changes in translational yield regulate tissue-specific expression of beta-glucuronidase. *Proc Natl Acad Sci USA* **84**: 9020-9024.

Bridges, C. D. (1976). Vitamin A and the role of the pigment epithelium during bleaching and regeneration of rhodopsin in the frog eye. *Exp Eye Res.* **22**(5): 435-55.

Briggs, C. E., D. Rucinski, P. J. Rosenfeld, T. Hirose, E. L. Berson and T. P. Dryja (2001). Mutations in ABCR (ABCA4) in patients with Stargardt macular degeneration or cone-rod degeneration. *Invest. Ophthalmol. Vis. Sci.* **42**(10): 2229–2236.

Buchanana, S. G. C. and N. J. Gay (1996). Structural and functional diversity in the leucine-rich repeat family of proteins. *Progress in Biophysics and Molecular Biology* **65**(1-2): 1-44.

Buraczynska, M., A. J. Mears, S. Zareparsi, R. Farjo, E. Filippova, Y. Yuan, S. P. MacNee, B. Hughes and A. Swaroop (2002). Gene expression profile of native human retinal pigment epithelium. *Invest Ophthalmol Vis Sci.* **43**(3): 603-7.

Burge, C. and S. Karlin (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**: 78-94.

Ciulla, T. A., R. P. Danis and A. Harris (1998). Age-related macular degeneration: a review of experimental treatments. *Surv. Ophthalmol* **43**: 134-146.

Collins, F. C. (1995). Positional cloning moves from perditional to traditional. *Nat Genet* **9**: 347-350.

Cruickshanks, K., R. Klein, B. Klein and D. Nondahl (2001). Sunlight and the 5-year incidence of early age-related maculopathy: the beaver dam eye study. *Archives Of Ophthalmology* **119**(2): 246-250.

Curcio, C. A. and C. L. Millican (1999). Basal linear deposit and large drusen are specific for early age-related maculopathy. *Arch Ophthalmol.* **117**(3): 329-39.

Daly, M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson and E. S. Lander (2001). High-resolution haplotype structure in the human genome. *Nat Genet* **29**: 229-232.

Dawson, E., G. R. Abecasis, S. Bumpstead, Y. Chen, S. Hunt, D. M. Beare, J. Pabilal and e. al (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**: 544-548.

De La Paz, M. A., M. A. Pericak-Vance, F. Lennon, J. L. Haines and J. M. Seddon (1997). Exclusion of TIMP3 as a candidate locus in age-related macular degeneration. *Invest Ophthalmol Vis Sci* **38**: 1060-1065.

den Hollander, A., M. van Driel, Y. de Kok, D. van de Pol, C. Hoyng, H. Brunner, A. Deutman and F. Cremers (1999). Isolation and mapping of novel candidate genes for retinal disorders using suppression subtractive hybridization. *Genomics* **58**: 240-249.

Diatchenko, L., A. Campbell, A. Chenchik, F. Moqadam, B. Huang, S. Lukyanov, K. Lukyanov, N. Gurskaya, E. D. Sverdlov and P. D. Siebert (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.

Diatchenko, L., S. Lukyanov, Y. Lau and P. Siebert (1999). Suppression subtractive hybridization: a versatile method for identifying differentially expressed genes. *Methods Enzymol* **303**: 349-380.

Diehn, M., G. Sherlock, G. Binkley, H. Jin, J. C. Matese, T. Hernandez-Boussard, C. A. Rees, J. M. Cherry, D. Botstein, P. O. Brown and A. A. Alizadeh (2003). SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucl. Acids. Res.* **31**(1): 219-23.

Dryja, T. P. (1997). Gene-based approach to human gene-phenotype correlations. *Proc. Natl. Acad. Sci. USA* **94**: 12117-12121.

Dryja, T. P., C. E. Briggs, E. L. Berson, P. J. Rosenfeld and M. Abitbol (1998). ABCR Gene and Age-Related Macular Degeneration. *Science* **279**: 1107 online http://www.sciencemag.org/cgi/reprint/279/5354/1107a.pdf.

Duguid, J. R. and M. C. Dinauer (1990). Library subtraction of in vitro cDNA libraries to identify differentially expressed genes in scrapie infection. *Nucleic Acids Res* **18**: 2789-92.

Dunaief, J. L., T. Dentchev, G. Ying and A. H. Milam (2002). The Role of Apoptosis in Age-Related Macular Degeneration. *Arch Ophthalmol.* **120**: 1435-1442.

EDCCS Group (1993). Antioxidant status and neovascular age-related macular degeneration. Eye Disease Case-Control Study Group. *Arch Ophthalmol.* **111**(1): 104-109.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics* **14**(9): 755-63.

Eldred, G. E. (1998). Lipofuscin and other storage deposits in the RPE. In: Marmor MF, Wolfensberger TJ, eds. The Retinal Pigment Epithelium: Function and Disease. Oxford, UK, Oxford University Press**:** 651-668.

Feeney-Burns, L., E. S. Hilderbrand and S. Eldrige (1984). Aging human RPE: morphometric analysis of macular, equatorial, and peripheral cells. *Invest Ophthalmol Vis Sci* **25**: 195-200.

Felbor, U., H. Schilling and B. H. Weber (1997). Adult vitelliform macular dystrophy is frequently associated with mutations in the peripherin/RDS gene. *Hum Mutation* **10**: 301-309.

Fickett, J. W. (1982). Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res* **10**: 5503-5518.

Freund, C. L., Q. L. Wang, S. Chen, B. L. Muskat, C. D. Wiles, V. C. Sheffield, S. G. Jacobson, R. R. McInnes, D. J. Zack and E. M. Stone (1998). De novo mutations in the CRX homeobox gene associated with Leber congenital amaurosis. *Nat Genet* **18**(4): 311-2.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly and D. Altshuler (2002). The structure of haplotype blocks in the human genome. *Science* **296**: 2225-2229.

Gieser, L. and A. Swaroop (1992). Expressed sequence tags and chromosomal localization of cDNA clones from a subtracted retinal pigment epithelium library. *Genomics* **13**: 873-6.

Green, W. R., P. J. McDonnell and J. H. Yeo (1985). Pathologic features of senile macular degeneration. *Ophthalmology* **92**: 615–627.

Guymer, R. H., R. McNeil, M. Cain, B. Tomlin, P. J. Allen, C. L. Dip and P. N. Baird (2002). Analysis of the Arg345Trp disease-associated allele of the EFEMP1 gene in individuals with early onset drusen or familial age-related macular degeneration. *Clin Experiment Ophthalmol.* **30**(6): 419-23.

Haeseleer, F., G.-F. Jang, Y. Imanishi, C. Driessen, M. Matsumura, P. Nelson and K. Palczewski (2002). Dual-substrate Specificity Short Chain Retinol Dehydrogenases from the Vertebrate Retina. *The Journal of Biological Chemistry* **277**(47): 45537-45546.

Hageman, G. S., L. V. Johnson, D. H. Anderson and R. F. Mullins (2001). An Integrated Hypothesis That Considers Drusen as Biomarkers of Immune-Mediated Processes at the RPE-Bruch's Membrane Interface in Aging and Age-Related Macular Degeneration. *Prog Retin Eye Res.* **20**(6): 705-732.

Hameed, A., S. Khaliq, M. Ismail, K. Anwar, N. D. Ebenezer, T. Jordan, S. Q. Mehdi, A. M. Payne and S. S. Bhattacharya (2000). A novel locus for Leber congenital amaurosis (LCA4) with anterior keratoconus mapping to chromosome 17p13. *Invest Ophthalmol Vis Sci* **41**(3): 629-33.

Hammond, B. R., B. R. Wooten and D. M. Snodderly (1996). Cigarette smoking and retinal carotenoids: implications for age-related macular degeneration. *Vision Res* **36**: 3003-3009.

Harman, D. (1981). The aging process. *Proc. Nat. Acad. Sci. USA* **78**: 7124-7128.

Harman, D. (1998). Aging phenomena and theories. *Ann. N. A. Acad. Sci.* **854**: 1-7.

He, S., H. M. Wang, T. E. Ogden and S. J. Ryan (1993). Transplantation of cultured human retinal pigment epithelium into rabbit subretina. *Graefe's Archive For Clinical And Experimental Ophthalmology* **231**(12): 737-742.

Höög, C. (1991). Isolation of a large number of novel mammalian genes by a differential cDNA library screening strategy. *Nucleic Acids Res* **17**: 6123-6127.

Huang, X. and A. Madan (1999). CAP3: a DNA sequence assembly program. *Genome Res* **9**: 868-877.

Hughes, B. A., R. P. Gallemore and S. S. Miller (1998). Transport mechanisms in the retinal pigment epithelium. In: Marmor MF, Wolfensberger TJ, eds. The Retinal Pigment Epithelium. Oxford, UK, Oxford University Press**:** 103-108.

Hutchinson, G. B. and M. R. Hayden (1992). The prediction of exons through an analysis of spliceable open reading frames. *Nucleic Acids Res.* **20**(13): 3453-62.

Huynen, M. A. and B. Snel (2000). Gene and context: integrative approaches to genome analysis. *Adv. Protein Chem.* **54**: 345-379.

Hyman, L., A. P. Schachat, Q. He and C. Leske (2000). Hypertension, Cardiovascular Disease, and Age-Related Macular Degeneration. *Arch Ophthalmol.* **118**(3).

Jinghua, T. C., E. Noriko, M. Kathryn, L. Yuanyuan, Z. Suiyuan, C. Christina, G. Barbara, R. Amir, M. Sally, S. Gail, A. C. Peter and J. Z. Donald (1999). Cloning and characterization of a secreted frizzled-related protein that is expressed by the retinal pigment epithelium. *Human Molecular Genetics*: 575-583.

Jobling, A. I., Z. Fang, D. Koleski and M. J. Tymms (2002). Expression of the ETS Transcription factor ELF3 in the Retina Pigment Epithelium. *Invest Ophthalmol Vis Sci* **43**(11): 3530-3537.

Joernvall, H., B. Persson, M. Krook, S. Atrian, R. Gonzalez-Duarte, J. Jeffery and D. Ghosh (1995). Short-chain dehydrogenases/reductases (SDR). *Biochemistry* **34**: 6003-6013.

Johnson, G. C., L. Esposito, B. Barratt, J., A. N. Smith, J. Heward, G. D. Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, R. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, C. Phillipa, E. Tuomilehto-Wolf, J. Tuomilehto, S. Gough, D. G. Clayton and J. A. Todd (2001). Haploype tagging for the identification of common disease genes. *Nat Genet* **29**: 233-237.

Jones, S. E., C. Jomary, J. Grist, H. J. Stewart and M. J. Neal (2000). Modulated expression of secreted frizzled-related proteins in human retinal degeneration. *Neuroreport* **11**(18): 3963-3967.

Katz, M. L., C. M. Drea and W. G. Robison (1986). Relationship between dietary retinol and lipofuscin in the retinal pigment epithelium. *Mech Age Dev*.

Klaver, C. C. W., J. M. Assink, A. Bergen and C. M. van Duijn (1998). ABCR Gene and Age-Related Macular Degeneration. *Science* **279**: 1107 online http://www.sciencemag.org/cgi/reprint/279/5354/1107a.pdf.

Klein, M. L., D. W. Schultz, A. Edwards, T. C. Matise, K. Rust, C. B. Berselli, K. Trzupek, R. G. Weleber, J. Ott, M. K. Wirtz and e. al. (1998). Age-related macular degeneration. Clinical features in a large family and linkage to chromosome 1q. *Arch. Ophthalmol.* **116**: 1082-1088.

Klein, R., B. E. Klein and K. L. P. Linton (1992). Prevalence of age-related maculopathy: The Beaver Dam Eye Study. *Ophthalmology* **99**: 933–943.

Klein, R., B. E. Klein, E. K. Marino, L. H. Kuller, C. Furberg, G. L. Burke and L. D. Hubbard (2003). Early age-related maculopathy in the cardiovascular health study. *Ophthalmology* **110**(1): 25-33.

Kohl, S., I. Giddings, D. Besch and Z. E. Apfelstedt-Sylla E, Wissinger B. (1998). The role of the peripherin/RDS gene in retinal dystrophies. *Acta Anat* **162**: 75-84.

Korf, I., P. Flicek, D. Duan and M. R. Brent (2001). Integrating genomic homology into gene structure prediction. *Bioinformatics* **17**(Suppl 1): S140-8.

Krämer, F., K. White, D. Pauleikhoff, A. Gehrig, L. Passmore, A. Rivera, G. Rudolph, U. Kellner, M. Andrassi, B. Lorenz, K. Rohrschneider, A. Blankenagel, B. Jurklies, H. Schilling, F. Schutt, F. G. Holz and B. H. Weber (2000). Mutations in the VMD2 gene are associated with juvenile-onset vitelliform macular dystrophy (Best disease) and adult vitelliform macular dystrophy but not age-related macular degeneration. *Eur J Hum Genet* **8**: 286-292.

Krzystolik, M. G., M. A. Afshari, A. P. Adamis, J. Gaudreault, E. S. Gragoudas, N. A. Michaud, W. Li, E. Connolly, C. A. O'Neill and J. W. Miller (2002). Prevention of experimental choroidal neovascularization with intravitreal anti-vascular endothelial growth factor antibody fragment. *Arch. Ophthalmol.* **120**: 338-346.

Lavail, M. M., L. Li, J. E. Turner and D. Yasumura (1992). Retinal pigment epithelial cell transplantation in RCS rats: normal metabolism in rescued photoreceptors. *Experimental Eye Research* **55**(4): 555-562.

Leibowitz, H. M., D. E. Krueger, L. R. Maunder, R. C. Milton, M. M. Kini, H. A. Kahn, R. J. Nickerson, J. Pool, T. L. Colton, J. P. Ganley, J. I. Loewenstein and T. R. Dawber (1980). The Framingham Eye Study monograph: An ophthalmological and epidemiological study of cataract, glaucoma, diabetic retinopathy, macular degeneration, and visual acuity in a general population of 2631 adults, 1973–1975. *Surv. Ophthalmol.* **24**: 335-610.

Liang, F. and B. Godley (2003). Oxidative stress-induced mitochondrial DNA damage in human retinal pigment epithelial cells: a possible mechanism for RPE aging and age-related macular degeneration. *Exp Eye Res* **76**(4): 397-403.

Liang, P. and A. B. Pardee (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**: 967-71.

Liew, C. C., D. M. Hwang, Y. W. Fung, C. Laurenssen, E. Cukerman, S. Tsui and C. Y. Lee (1994). A catalogue of genes in the cardiovascular system as identified by expressed sequence tags. *Proc Natl Acad Sci USA* **91**: 10645-9.

Marlhens, F., C. Bareil, J. M. Griffoin, E. Zrenner, P. Amalric, C. Eliaou, S. Y. Liu, E. Harris, T. M. Redmond, B. Arnaud, M. Claustres and C. P. Hamel (1997). Mutations in RPE65 cause Leber's congenital amaurosis. *Nat Genet* **17**(2): 139-41.

Marmor, M. F. (1998). Structure, function and disease of retinal epithelium. In: Marmor MF, Wolfensberger TJ, eds. The Retinal Pigment Epithelium: Function and Disease. Oxford, NY:, Oxford University Press**:** 3-9.

Marquardt, A., H. Stöhr, L. Passmore, F. Kramer, A. Rivera and B. Weber (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**: 1517-1525.

Maw, M. A., B. Kennedy, A. Knight, R. Bridges, K. E. Roth, E. J. Mani, J. K. Mukkadan, D. Nancarrow, J. W. Crabb and M. J. Denton (1997). Mutation of the gene encoding cellular retinaldehyde-binding protein in autosomal recessive retinitis pigmentosa. *Nat Genet.* **17**: 198-200.

McClelland, M., F. Mathieu-Daude and J. Welsh (1995). RNA fingerprinting and differential display using arbitrarily primed PCR. *Trends Genet* **11**: 242-6.

McKusick, V. A. (1998). Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders, 12th edn., The Johns Hopkins University Press, Baltimore, MD.

Meldrum, M., M. Lejk and P. Guy (1993). SSADM Techniques: An introduction to Version 4, Chartwell-Bratt, Lund, Sweden.

Meyers, S. M., T. Greene and F. A. Gutman (1995). A twin study of age-related macular degeneration. *Am J Ophthalmol* **120**: 757-66.

Miller, R. D., J. W. Hoffmann, P. P. Powell, J. W. Kyle, J. M. Shipley, D. R. Bachinsky and W. S. Sly (1990). Cloning and characterization of the human beta-glucuronidase gene. *Genomics* **7**: 280-283.

O'Donovan, M. C., P. J. Oefner, S. C. Roberts, J. Austin, B. Hoogendoorn and C. e. a. Guy (1998). Blind analysis of denaturing high-performance liquid chromatography as a tool for mutation detection. *Genomics* **52**: 44-9.

Ohno-Matsui, K., I. Morita, J. Tombran-Tink, D. Mrazek, M. Onodera, T. Uetama, M. Hayano, S. I. Murota and M. Mochizuki (2001). Novel mechanism for age-related macular degeneration: an equilibrium shift between the angiogenesis factors VEGF and PEDF. *J Cell Physiol.* **189**(3): 323-33.

Ohno-Matsui, K., T. Yoshida, T. Uetama, M. Mochizuki and I. Morita (2003). Vascular endothelial growth factor upregulates pigment epithelium-derived factor expression via VEGFR-1 in human retinal pigment epithelial cells. *Biochem Biophys Res Commun.* **303**(3): 962-7.

Ophoff, R. A., G. M. Terwindt, M. N. Vergouwe, R. van Eijk, P. Oefner, S. M. Hoffman, J. E. Lamerdin, H. W. Mohrenweiser, D. E. Bulman, M. Ferrari, J. Haan, D. Lindhout, G. J. van Ommen, M. H. Hofker, M. D. Ferrari and R. R. Frants (1996). Familial hemiplegic migraine and episodic ataxia type-2 are caused by mutations in the Ca2+ channel gene CACNL1A4. *Cell* **87**(3): 543-52.

Orita, M., H. Iwahana, H. Kanazawa, K. Hayashi and T. Sekiya (1989). Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc Natl Acad Sci U S A* **86**(8): 2766-70.

Parra , G., P. Agarwal, J. F. Abril , T. Wiehe, J. W. Fickett and R. c. Guigó (2003). Comparative Gene Prediction in Human and Mouse. *Genome Res* **13**(1): 108-117.

Patil, N., A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. N. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. A. Fodor and D. R. Cox (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719-1723.

Pearson, W. R. and D. J. Lipman (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85**: 2444–2448.

Perrault, I., J. M. Rozet, P. Calvas, S. Gerber, A. Camuzat, H. Dollfus, S. Chatelin, E. Souied, I. Ghazi, C. Leowski, M. Bonnemaison, D. Le Paslier, J. Frezal, J. L. Dufier, S. Pittler, A. Munnich and J. Kaplan (1996). Retinal-specific guanylate cyclase gene mutations in Leber's congenital amaurosis. *Nat Genet.* **14**(4): 461-4.

Petrukhin, K., M. Koisti, B. Bakall, W. Li, G. Xie, T. Marknell, O. Sandgren and e. al (1998). Identification of the gene responsible for Best macular dystrophy. *Nat Genet* **19**: 241-247.

Presta, L. G., H. Chen, S. J. O'Connor, V. Chisholm, Y. G. Meng, L. Krummen, M. Winkler and N. Ferrara (1997). Humanization of an anti-vascular endothelial growth factor monoclonal antibody for the therapy of solid tumors and other disorders. *Cancer Res.* **57**(20): 4593-9.

Pruitt, K. and D. Maglott (2001). RefSeq and LocusLink : NCBI gene-centered resources. *Nucleic Acids Res.* **29**: 137-140.

Rando, R. R. (1996). Polyenes and vision. *Chem Biol.* **3**: 255-262.

Risch, N. and K. Merikangas (1996). The future of genetic studies of complex human diseases. *Science* **273**: 1516-1517.

Rivera, A., W. K., H. Stöhr, K. Steiner, N. Hemmrich, T. Grimm, B. Jurklies, B. Lorenz, H. P. Scholl, E. Apfelstedt-Sylla and B. H. Weber (2000). A comprehensive survey of sequence variation in the ABCA4 (ABCR) gene in Stargardt disease and age-related macular degeneration. *Am J Hum Genet* **67**: 800-813.

Rommens, J. M., M. C. Iannuzzi, B. Kerem and e. al. (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245**: 1059-1065.

Rychlik, W. and R. Rhoads (1989). A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. *Nucl. Acids. Res.* **17**(21): 8543-51.

Saari, J. C. (2000). Biochemistry of Visual Pigment Regeneration The Friedenwald Lecture. *IOVS* **41**: 337-348.

Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuche, G. T. Horn, K. M. Mullis and H. A. Erlich (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**: 487-491.

Sarks, S. H., D. van Driel, L. Maxwell and M. Killingsworth (1980). Softening of drusen and subretinal neovascularization. *Trans Ophthalmol Soc U K.* **100**: 414-422.

Sarna, T., J. Burkeb, W. Korytowskia, M. Róanowskaa, C. Skumatzb, A. Zarbaa and M. Zarbaa (2003). Loss of melanin from human RPE with aging: possible role of melanin photooxidation. *Exp Eye Res* **76**(1): 89-98.

Satoh, K., M. Hata and H. Yokota (2002). A novel member of the leucine-rich repeat superfamily induced in rat astrocytes by beta-amyloid. *Biochem Biophys Res Commun.* **290**(2): 756-62.

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

Seddon, J. M., U. A. Ajani and B. D. Mitchell (1997). Familial aggregation of age-related maculopathy. *Am J Ophthalmol* **123**: 199-206.

Shamsi, F. A. and M. Boulton (2001). Inhibition of RPE lysosomal and antioxidant activity by the age pigment lipofuscin. *Invest Ophthalmol Vis Sci* **42**: 3041-3046.

Sharma, S., C. J. T., D. N.G., P. A. Campochiaro and D. J. Zack (2002). Identification of novel bovine RPE and retinal genes by subtractive hybridization. *Mol Vis* **8**: 251-258.

Sharon, D., S. Blackshaw, C. L. Cepko and T. P. Dryja (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 U S A.* **99**(1): 315-20.

Shastry, B. S. and M. T. Trese (1999). Evaluation of the peripherin/RDS gene as a candidate gene in families with age-related macular degeneration. *Ophthalmologica* **213**: 165-170.

Sheng, Y., P. Gouras, H. Cao, L. Berglin, H. Kjeldbye, R. Lopez and H. Rosskothen (1995). Patch transplants of human fetal retinal pigment epithelium in rabbit and monkey retina. *Invest Ophthalmol Vis Sci* **36**(2): 381-90.

Shepherd, J. C. W. (1981). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc Natl Acad Sci USA* **78**: 1596-1600.

Spraul, C. W., G. E. Lang and H. E. Grossniklaus (1996). Morphometric analysis of the choroid, Bruch's membrane, and retinal pigment epithelium in eyes with age-related macular degeneration. *Invest. Ophthalmol. Vis. Sci.* **37**: 2724–2735.

Srivastava, S., A. Chandra, A. Bhatnagar, S. K. Srivastava and N. H. Ansari (1995). Lipid peroxidation product, 4-hydroxynonenal and its conjugate with GSH are excellent substrates of bovine lens aldose reductase. *Biochem. Biophys. Res. Commun.* **217**: 741-746.

Staden, R. and A. D. McLachlan (1982). Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res* **10**: 141-156.

Stockton, D. W., R. A. Lewis, E. B. Abboud, A. Al-Rajhi, M. Jabak, K. L. Anderson and J. R. Lupski (1998). A novel locus for Leber congenital amaurosis on chromosome 14q24. *Hum Genet.* **103**(3): 328-33.

Stoesser, G., W. Baker, A. van den Broek, M. Garcia-Pastor, C. Kanz, T. Kulikova, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, F. Nardone, P. Stoehr, M. A. Tuli, K. Tzouvara and R. Vaughan (2003). The EMBL Nucleotide Sequence Database: major new developments. *Nucl. Acids. Res.* **31**(1): 17-22.

Stollberg, J., J. Urschitz, Z. Urban and C. D. Boyd (2000). A Quantitative Evaluation of SAGE. *Genome Res.* **10**(8): 1241-1248.

Stone, E. M., A. J. Lotery, F. L. Munier, E. Heon, B. Piguet, R. H. Guymer, K. Vandenburgh, P. Cousin, D. Nishimura, R. Swiderski, G. Silvestri, D. Mackey, G. S. Hageman, A. C. Bird, V. C. Sheffield and D. F. Schorderet (1999). A single EFEMP1 mutation associated with both Malattia Leventinese and Doyne honeycomb retinal dystrophy. *Nat Genet* **22**(2): 199-202.

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

Swaroop, A. and D. J. Zack (2002). Transcriptome analysis of the retina. *Genome Biology* **3**(8): 1022.1-2022.4.

Tateno, Y., Y. Imanishi, S. Miyazaki, K. Fukami-Kobayashi, N. Saitou, H. Sugawara and T. Gojobori (2002). DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucl. Acids. Res.* **30**(1): 27-30.

Tatusova, T. A. and T. L. Madden (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174**: 247-250.

Thumann, G. (2001). Development and Cellular Functions of the Iris Pigment Epithelium. *Survey of Ophthalmology* **45**(4): 345-354.

Thumann, G., N. Kociok, K. U. Bartz-Schmidt, P. Esser, U. Schraermeyer and K. Heimann (1999). Detection of mRNA for proteins involved in retinol metabolism in iris pigment epithelium. *Graefe's Archive for Clinical and Experimental Ophthalmology* **237**(12): 1046-1051.

Uberbacher, E. C. and R. J. Mural (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc Natl Acad Sci U S A.* **88**(24): 11261-5.

Van Eerdewegh, P., R. D. Little, J. Dupuis, R. G. Del Mastro, K. Falls, J. Simon, D. Torrey, S. Pandit, J. McKenny, K. Braunschweiger, A. Walsh, Z. Liu, B. Hayward, C. Folz, S. P. Manning, A. Bawa, L. Saracino, M. Thackston, Y. Benchekroun, N. Capparell, M. Wang, R. Adair, Y. Feng, J. Dubois, M. G. FitzGerald, H. Huang, R. Gibson, K. M. Allen, A. Pedan, M. R. Danzig, S. P. Umland, R. W. Egan, F. M. Cuss, S. Rorke, J. B. Clough, J. W. Holloway, S. T. Holgate and K. T. P. (2002). Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature* **418**(6896): 426-430.

Velculescu, V. E., L. Zhang, B. Vogelstein and K. W. Kinzler (1995). Serial analysis of gene expression. *Science* **270**: 484-7.

Verdugo, M. E. and J. Ray (1997). Age-related increased in activity of specific lysosomal enzymes in the Human retinal pigment epithelium. *Exp Eye Res.* **65**: 231-240.

Wang, D. G., J. B. Fan, C. J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour and e. al (1998). Large Scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077-1082.

Weber, B. H., G. Vogt, R. C. Pruett, H. Stöhr and U. Felbor (1994). Mutation in the tissue inhibitor of metalloproteinases-3 (TIMP3). *Nat Genet* **8**: 352-356.

Webster, A., E. Heon, A. Lotery, K. Vandenburgh, T. Casavant, K. Oh, G. Beck, G. Fishman, B. Lam, A. Levin, J. Heckenlively, S. Jacobson, R. Weleber, V. Sheffield and E. Stone (2001). An analysis of allelic variation in the ABCA4 gene. *Invest Ophthalmol Vis Sci* **42**: 1179-1189.

Wistow, G., S. L. Bernstein, M. K. Wyatt, R. N. Fariss, A. Behal, J. W. Touchman, G. Bouffard, D. Smith and K. Peterson (2002a). Expressed sequence tag analysis of human RPE/choroid for the NEIBank Project: Over 6000 non-redundant transcripts, novel genes and splice variants. *Molecular Vision* **8**: 205-220.

Wistow, G., S. L. Bernstein, M. K. Wyatt, S. Ray, A. Behal, J. W. Touchman, G. Bouffard, D. Smith and K. Peterson (2002b). 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.

Wistow, G. (2002c). A project for ocular bioinformatics: NEIBank. *Mol Vis* **15**: 161-163.

Wyllie, A. H., F. F. R. Kerr and A. R. Currie (1980). Cell death: the significance of apoptosis. *Int Rev Cytol.* **68**: 251-306.

Xu, G. Z., W. W. Li and T. M. O. (1996). Apoptosis in human retinal degenerations. *Trans Am Ophthalmol Soc.* **94**: 411-430.

Xu, Y., J. R. Einstein, R. J. Mural, M. Shah and E. C. Uberbacher (1994). An improved system for exon recognition and gene modeling in human DNA sequences. *In Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 376-384*.

Yamamoto, H., A. Simon, U. Eriksson, E. Harris, E. L. Berson and T. P. Dryja (1999). Mutations in gene encoding 11-cis retinol dehydrogenase cause delayed dark adaptation and fundus albipunctatus. *Nat Genet.* **22**: 188-191.

Yeh, R., L. P. Lim and C. B. Burge (2001). Computational Inference of Homologous Gene Structures in the Human Genome. *Genome Res* **11**(5): 803-816.

Yoshinaka, T., K. Nishii, K. Yamada, H. Sawada, E. Nishiwaki, K. Smith, K. Yoshino, H. Ishiguro and S. Higashiyama (2002). Identification and characterization of novel mouse and human ADAM33s with potential metalloprotease activity. *Gene* **282**(1-2): 227236.

Zarbin, M. A. (1998). Age-related macular degeneration: review of pathogenesis. *Eur. J. Ophthalmol.* **8**: 199-206.

Zhang, L., W. Zhou, V. E. Velculescu, S. E. Kern, R. H. Hruban, S. R. Hamilton, B. Vogelstein and K. W. Kinzler (1997). Gene Expression Profiles in Normal and Cancer Cells. *Science* **276**(5316): 1268-72.

APPENDIX

1. ESTs representing known genes in the SSRPE cDNA library classified by function

Gene Name (Gene symbol)	Accession Number	Frequency
Cell defence		
ATX1 antioxidant protein 1 homolog (yeast) (ATOX1)	NM_004045	1
Glutathione S-transferase M5 (GSTM5)	NM_000851	2
Glutathione S-transferase M (GSTM)	NM_000850	44
BCL2-like 1 (BCL2L1)	NM_138578	3
Eukaryotic translation elongation factor 1 epsilon 1(EEF1E1)	NM_004280	1
Monoamine oxidase B MAOB	NM_000898	18
BCL2/adenovirus E1B 19kDa interacting protein 2 (BNIP2)	NM_004330	1
Microsomal glutathione S-transferase 1 (MGST1)	NM_145791	3
Melanoma antigen, family D, 1 (MAGED1)	NM_006986	1
Forkhead box O1A (rhabdomyosarcoma) (FOXO1A)	NM_002015	1
Requiem, apoptosis response zinc finger gene (REQ)	NM_006268	1
Glutathione peroxidase 4 (phospholipid hydroperoxidase) (GPX4)	NM_002085	3
Defender against cell death 1 (DAD1)	NM_001344	1
Testis enhanced gene transcript (BAX inhibitor 1) (TEGT)	NM_003217	2
Apoptosis related protein (APR-3)	NM_016085	5
Glucose phosphate isomerase (GPI)	NM_000175	2
Heat shock 90kDa protein 1, beta (HSPCB)	NM_007355	1
Peroxiredoxin 1 (PRDX1)	NM_002574	3
Peroxiredoxin 2 (PRDX2)	NM_005809	4
Clusterin (CLU)	NM_001831	6
Cell division		
Likely ortholog of mouse septin 8 (SEPT8)	XM_034872	1
Cyclin-dependent kinase 2 (CDK2)	NM_001798	1
LAG1 longevity assurance homolog 2 (S. cerevisiae) (LASS2)	NM_013384	1
Growth differentiation factor 11 (GDF11)	NM_005811	1
Cell structure/maintenance		
Centaurin, delta 2 (CENTD2)	NM_139181	1
Component of oligomeric golgi complex 5 (COG5)	NM_006348	1
Component of oligomeric golgi complex 7 (COG7)	NM_153603	1
Ectonucleotide pyrophosphatase/phosphodiesterase 2 (ENPP2)	NM_006209	2
Ferritin, heavy polypeptide 1 (FTH1)	NM_002032	4
Skeletal muscle and kidney enriched inositol phosphatase (SKIP)	NM_016532	15
Solute carrier family 9 (sodium/hydrogen exchanger), isoform 3 regulatory factor 1 (SLC9A3R1)	NM_004252	1
CLIP-170-related protein (CLIPR-59)	NM_015526	1
Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin) (SPTAN1)	NM_003127	1
Histone 2, H2be (HIST2H2BE)	NM_003528	1
Serine (or cysteine) proteinase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 1 (SERPINF1)	NM_002615	1
Dynein light chain 2 (Dlc2)	NM_080677	1
Myeloid leukemia factor 2 (MLF2)	NM_005439	1
Myosin, light polypeptide 6, alkali, smooth muscle and non-muscle (MYL6)	NM_079425	2
Tubulin, gamma 1 (TUBG1)	NM_001070	1
Protease, serine, 11 (IGF binding) (PRSS11)	NM_002775	6
Histone 1, H2bd (HIST1H2BD)	NM_138720	1
Cell-cell signalling		
Sparc/osteonectin, cwcv and kazal-like domains proteoglycan ((SPOCK)	NM_004598	6
V-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (SRC)	NM_005417	2
Secreted frizzled-related protein 5 (SFRP5)	NM_003015	1
Transforming growth factor, beta-induced, 68kDa (TGFBI)	NM_000358	1
Coagulation factor II (thrombin) receptor-like 2 (F2RL2)	NM_004101	1
FLN29 gene product (FLN29)	NM_006700	2
Flotillin 2 (FLOT2)	NM_004475	1
Gamma-aminobutyric acid (GABA) receptor, rho 2 (GABRR2)	NM_002043	2
Gap junction protein, beta 1, 32kDa (connexin 32, Charcot-Marie-Tooth neuropathy, X-linked) (GJB1)	NM_000166	1
Shroom-related protein (ShrmL)	NM_020859	1
Dual specificity phosphatase 6 (DUSP6)	NM_001946	1
Secreted frizzled-related protein 2 (SFRP2)	XM_050625	4
MT-protocadherin (KIAA1775)	NM_033100	1
RYK receptor-like tyrosine kinase (RYK)	NM_002958	1
Retinal degeneration, slow (retinitis pigmentosa 7) (RDS)	NM_000322	10
Natriuretic peptide receptor B/guanylate cyclase B (NPR2)	NM_000907	1
protocadherin gamma subfamily A, 5 (PCDHG@)	NM_002588	1
Pygopus 2 (PYGO2)	NM_138300	1
Seven in absentia homolog 2 (Drosophila) (SIAH2)	NM_005067	1
Ras homolog gene family, member C (ARHC)	NM_005167	1

1. (continued)

Gene Name (Gene symbol)	Accession Number	Frequency
Transmembrane 4 superfamily member 2 (TM4SF2)	NM_004615	5
Syndecan 4 (amphiglycan, ryudocan) (SDC4)	NM_002999	1
Cadherin 3, type 1, P-cadherin (placental) (CDH3)	NM_001793	68
Tetraspan 3 (TSPAN-3)	NM_005724	5
Ortholog of mouse mitogen activated protein kinase binding protein 1 (MAPKBP1)	XM_031706	1
Basigin (OK blood group) (BSG)	NM_001728	1
Mitogen-activated protein kinase kinase 1 (MAP2K1)	NM_002755	1
G protein-coupled receptor, family C, group 5, member B (GPRC5B)	NM_016235	3
Colony stimulating factor 1 receptor, formerly McDonough feline sarcoma viral (v-fms) oncogene homolog (CSF1R)	NM_005211	2
Guanine nucleotide binding protein (G protein), beta polypeptide 3 (GNB3)	NM_002075	1
Guanine nucleotide binding protein (G protein), alpha transducing activity polypeptide 1 (GNAT1)	NM_144499	5
Vascular endothelial growth factor (VEGF)	NM_003376	2
Active BCR-related gene (ABR)	NM_021962	4
Guanine nucleotide binding protein (G protein), gamma transducing activity polypeptide 1 (GNGT1)	NM_021955	3
Development		
Bone morphogenetic protein 7 (osteogenic protein 1) (BMP7)	NM_001719	2
Protein tyrosine phosphatase-like member a (PTPLA	NM_014241	1
Frizzled-related protein (FRZB)	NM_001463	7
Energy metabolism		
Acetyl-Coenzyme A synthetase 2 (ADP forming) (ACAS2)	NM_018677	1
Creatine kinase, mitochondrial 1 (ubiquitous) (CKMT1)	NM_020990	5
Cytochrome c oxidase subunit IV isoform 1 (COX4I1)	NM_001861	1
ytochrome c oxidase subunit VIIc (COX7C)	NM_001867	1
ATP synthase, H+ transporting, mitochondrial F0 complex, subunit c (subunit 9), isoform 2 (ATP5G2)	NM_005176	1
ATP synthase, H+ transporting, mitochondrial F1 complex, beta polypeptide (ATP5B)	NM_001686	1
ATP synthase, H+ transporting, mitochondrial F0 complex, subunit g (ATP5L)	NM_006476	1
Similar to citrate synthase precursor; Citrate synthase, mitochondrial (LOC284438)	XM_209202	1
Optic atrophy 3 (autosomal recessive, with chorea and spastic paraplegia) (OPA3)	NM_025136	1
(LOC136234)		1
Cytochrome c oxidase subunit VIII (COX8)	NM_004074	2
Acetyl-Coenzyme A acyltransferase 2 (ACAA2)	NM_006111	1
Extracellular matrix maintenance		
procollagen C-endopeptidase enhancer 2 (PCOLCE2)	NM_013363	1
Tissue inhibitor of metalloproteinase 2 (TIMP2)	NM_003255	2
Tissue inhibitor of metalloproteinase 3 (Sorsby fundus dystrophy, pseudoinflammatory) (TIMP3)	NM_000362	5
EGF-containing fibulin-like extracellular matrix protein 1(EFEMP1)	NM_004105	2
Fibromodulin (FMOD)	NM_002023	18
Lysosomal enzyme		
Cathepsin K (pycnodysostosis) (CTSK)	NM_000396	1
Sulfatase 1 (SULF1)	NM_015170	1
Metabolism		
like-glycosyltransferase (LARGE)	NM_004737	3
Protein phosphatase 1B (formerly 2C), magnesium-dependent, beta isoform (PPM1B)	NM_002706	1
Phospholipase A2, group V (PLA2G5)	NM_000929	2
Triosephosphate isomerase 1 (TPI1)	NM_000365	3
Thiamin pyrophosphokinase 1 (TPK1)	NM_022445	1
Lactate dehydrogenase B; (LDHB)	NM_002300	1
UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 1 (B4GALT1)	NM_001497	1
Phosphatidylserine synthase 1 (PTDSS1)	NM_014754	1
Tyrosinase-related protein 1 (TYRP1)	NM_000550	2
Pipecolic acid oxidase (PIPOX)	NM_016518	8
Microsomal NAD+-dependent retinol dehydrogenase 4 (RODH-4)	NM_003708	1
Phosphoribosyl pyrophosphate synthetase-associated protein 1 (PRPSAP1)	NM_002766	1
Phospholipase A2, group VII (platelet-activating factor acetylhydrolase, plasma) (PLA2G7)	NM_005084	3
Plasma glutamate carboxypeptidase (PGCP)	NM_016134	2

1. (continued)

Gene Name (Gene symbol)	Accession Number	Frequency
Peroxisomal biogenesis factor 11B (PEX11B)	NM_003846	1
Pyruvate dehydrogenase kinase, isoenzyme 2 (PDK2)	NM_002611	8
Isocitrate dehydrogenase 3 (NAD+) beta (IDH3B)	NM_174856	3
ADP-ribosyltransferase 3 (ART3)	NM_001179	1
Aldo-keto reductase family 1, member A1 (aldehyde reductase) (AKR1A1)	NM_006066	1
Dehydrogenase/reductase (SDR family) member 4 (DHRS4)	NM_021004	3
Isocitrate dehydrogenase 3 (NAD+) gamma (IDH3G)	NM_174869	2
Elongation of very long chain fatty acids (FEN1/Elo2, SUR4/Elo3, yeast)-like 2 (ELOVL2)	NM_017770	3
Elongation of very long chain fatty acids (FEN1/Elo2, SUR4/Elo3, yeast)-like 1 (ELOVL1)	NM_022821	7
Sphingomyelin phosphodiesterase 1, acid lysosomal (acid sphingomyelinase) (SMPD1)	NM_000543	2
Pleckstrin homology domain containing, family A (phosphoinositide binding specific) member 1 (PLEKHA1)	NM_021622	1
Cystatin C (amyloid angiopathy and cerebral hemorrhage) (CST3)	NM_000099	19
Enhancer of rudimentary homolog (Drosophila) (ERH)	NM_004450	2
Glutamic-oxaloacetic transaminase 2, mitochondrial (aspartate aminotransferase 2) (GOT2)	NM_002080	1
Fatty acid desaturase 1 (FADS1)	NM_013402	1
Deoxyhypusine synthase (DHPS)	NM_001930	1
Sulfite oxidase (SUOX)	NM_000456	2
Dolichyl-diphosphooligosaccharide-protein glycosyltransferase (DDOST)	NM_005216	1
Dimerization cofactor of hepatocyte nuclear factor 1 (HNF1) from muscle (DCOHM)	NM_032151	1
Glutamic-oxaloacetic transaminase 1, soluble (aspartate aminotransferase 1) (GOT1)	NM_002079	1
Glutamate-ammonia ligase (glutamine synthase) (GLUL)	NM_002065	5
Dolichyl-phosphate mannosyltransferase polypeptide 2, regulatory subunit (DPM2)	NM_152690	2
Carbonic anhydrase XII (CA12)	NM_001218	1
Hydroxyacyl-Coenzyme A dehydrogenase, type II (HADH2)	NM_004493	1
Succinate dehydrogenase complex, subunit B, iron sulfur (Ip) (SDHB)	NM_003000	5
Vitamin A responsive; cytoskeleton related (WA)	NM_006407	3
Hydroxysteroid (17-beta) dehydrogenase 8 (HSD17B8)	NM_014234	1
Isovaleryl Coenzyme A dehydrogenase (IVD)	NM_002225	2
Emopamil binding protein (sterol isomerase) (EBP)	NM_006579	1
Succinate dehydrogenase complex, subunit A, flavoprotein (Fp) (SDHA)	NM_004168	5
Glyceraldehyde-3-phosphate dehydrogenase (GAPD)	NM_002046	5
Carbonic anhydrase XI (CA11)	NM_001217	2
Carbonic anhydrase XIV (CA14)	NM_012113	27
Spermidine/spermine N1-acetyltransferase (SAT)	NM_002970	1
Solute carrier family 20 (phosphate transporter), member 1 (SLC20A1)	NM_005415	2
Ceroid-lipofuscinosis, neuronal 2, late infantile (Jansky-Bielschowsky disease) (CLN2)	NM_000391	3
Stromal cell-derived factor 2 (SDF2)	NM_006923	2
Neuron growth and function		
Olfactomedin 1 (OLFM1)	NM_014279	1
Putative small membrane protein (NID67)	NM_032947	1
Erythrocyte membrane protein band 4.1-like 1 (EPB41L1)	XM_047295	2
Neuronal proteinv(NP25)	NM_013259	1
Glycoprotein M6A (GPM6A)	NM_005277	1
Glycoprotein M6B (GPM6B)	NM_005278	12
Spastic ataxia of Charlevoix-Saguenay (sacsin) (SACS)	NM_014363	1
Phagocytosis		
c-mer proto-oncogene tyrosine kinase (MERTK)	NM_006343	1
Phototransduction		
rhodopsin (opsin 2, rod pigment) (retinitis pigmentosa 4, autosomal dominant) (RHO)	NM_000539	37
unc-119 homolog (C. elegans) (UNC119)	NM_054035	7
Sine oculis homeobox homolog 3 (Drosophila) (SIX3)	NM_005413	5
S-antigen; retina and pineal gland (arrestin) (SAG)	NM_000541	1
Recoverin (RCV1)	NM_002903	18
Pleckstrin homology domain containing, family B (evectins) member 1 (PLEKHB1)	NM_021200	5
ATP-binding cassette, sub-family A (ABC1), member 4 (ABCA4)	NM_000350	3

1. (continued)

Gene Name (Gene symbol)	Accession Number	Frequency
Protein degradation		
protein degradation; homocysteine-inducible, endoplasmic reticulum stress-inducible, ubiquitin-like domain member 1 (HERPUD1)	NM_014685	1
Protein degradation; SMT3 suppressor of mif two 3 homolog 1 (yeast) (SMT3H1)	NM_006936	1
Ubiquitin-activating enzyme E1 (A1S9T and BN75 temperature sensitivity complementing) (UBE1)	NM_003334	1
26S proteasome-associated pad1 homolog (POH1)	NM_005805	1
Ubiquitin-conjugating enzyme E2M (UBC12 homolog, yeast) (UBE2M)	NM_003969	1
Ring finger protein 25 (RNF25)	NM_022453	1
Vitamin A metabolism/transport		
Lecithin retinol acyltransferase (LRAT)	NM_004744	33
Retinal pigment epithelium-specific protein 65kDa (RPE65)	NM_000329	52
Retinal G protein coupled receptor (RGR)	NM_002921	37
Beta-carotene 15, 15'-dioxygenase (BCDO1)	NM_017429	13
Retinol dehydrogenase 5 (11-cis and 9-cis) (RDH5)	NM_002905	1
Retinol dehydrogenase 12 (all-trans and 9-cis) (RDH12)	NM_152443	5
Retinol dehydrogenase 10 (all-trans) (RDH10)	NM_172037	35
Retinol binding protein 4, plasma (RBP4)	NM_006744	1
Retinol binding protein 1, cellular (RBP1)	NM_002899	7
Retinaldehyde binding protein 1 (RLBP1)	NM_000326	63
Transcription factor		
Eukaryotic translation initiation factor 3, subunit 2 beta, 36kDa (EIF3S2)	NM_003757	1
Orthodenticle homolog 2 (Drosophila) (OTX2)	NM_021728	3
Sex comb on midleg homolog 1 (Drosophila) (SCMH1)	NM_012236	2
Fragile X mental retardation 2 (FMR2)	NM_002025	1
Basic helix-loop-helix domain containing, class B, 2 (BHLHB2)	NM_003670	1
Trinucleotide repeat containing 15 (TNRC15)	XM_209467	78
Enhancer of zeste homolog 1 (Drosophila) (EZH1)	NM_001991	1
Pilin-like transcription factor (PILB)	NM_012228	1
AE binding protein 1 (AEBP1)	NM_001129	1
Likely ortholog of mouse gene trap locus 3 (GTL3)	NM_013242	2
Transforming growth factor beta-stimulated protein (TSC22)	NM_006022	1
Putative DNA/chromatin binding motif (PLU-1)	NM_006618	1
SRB7 suppressor of RNA polymerase B homolog (yeast) (SURB7)	NM_004264	1
Neural retina leucine zipper (NRL)	NM_006177	2
Transport		
Solute carrier family 16 member 1 (SLC16A1)	NM_003051	1
Coatomer protein complex, subunit alpha (COPA)	NM_004371	1
Solute carrier family 13, member 3 (SLC13A3)	NM_022829	2
Solute carrier family 21, member 14 (SLC21A14)	NM_017435	1
Solute carrier family 22 (organic anion transporter), member 8 (SLC22A8)	NM_004254	2
Solute carrier family 24, member 1 (SLC24A1)	NM_004727	1
RAB7, member RAS oncogene family (RAB7)	NM_004637	1
Solute carrier family 1 (glutamate transporter), member 7 (SLC1A7)	NM_006671	5
Low density lipoprotein-related protein 1B (deleted in tumors) (LRP1B)	NM_018557	1
Solute carrier family 25 (mitochondrial carrier; oxoglutarate carrier), member 11 (SLC25A11)	NM_003562	1
Solute carrier family 2 (facilitated glucose transporter), member 1 (SLC2A1)	NM_006516	35
X transporter protein 3 (XT3)	NM_020208	36
Solute carrier family 41, member 1 (SLC41A1)	NM_173854	3
Serum/glucocorticoid regulated kinase; (SGK)	NM_005627	1
Solute carrier family 4, sodium bicarbonate cotransporter, member 5 (SLC4A5)	NM_033323	22
Solute carrier family 6, member 13(SLC6A13)	NM_016615	8
RAB5B, member RAS oncogene family (RAB5B)	NM_002868	1
Sorting nexin 5 (SNX5)	NM_152227	2
Member RAS onocogene family (RAB15)	XM_085123	1
Phosphotidylinositol transfer protein (PITPN)	NM_006224	1
ATP-binding cassette, sub-family C (CFTR/MRP), member 5 (ABCC5)	NM_005688	4
Syntaxin binding protein 1 (STXBP1)	NM_003165	2
Solute carrier family 38, member 3 (SLC38A3)	NM_006841	1
ATPase, H+ transporting, lysosomal V0 subunit a isoform 1 (ATP6V0A1)	NM_005177	8
Myosin VIIA and Rab interacting protein (MYRIP)	NM_015460	1
LAT1-3TM protein (LAT1-3TM)	NM_031211	1
Myosin VA (heavy polypeptide 12, myoxin) (MYO5A)	NM_000259	1
Transmembrane 9 superfamily member 1 (TM9SF1)	NM_006405	1
ATP-binding cassette, sub-family G (WHITE), member 2 (ABCG2)	NM_004827	1
Importin 13 (IPO13)	NM_014652	1

1. (continued)

Gene Name (Gene symbol)	Accession Number	Frequency
Similar to Na-Ca exchanger 5 [Mus musculus] (LOC283652)	XM_208771	3
NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 4, 9kDa (NDUFA4)	NM_002489	1
Transient receptor potential cation channel, subfamily M, member 1 (TRPM1)	NM_002420	13
Transient receptor potential cation channel, subfamily M, member 3 (TRPM3)	NM_020952	79
ATPase, Na+/K+ transporting, beta 2 polypeptide (ATP1B2)	NM_001678	17
ATPase, Na+/K+ transporting, alpha 3 polypeptide (ATP1A3)	NM_152296	2
Transthyretin (prealbumin, amyloidosis type I) (TTR)	NM_000371	87
ADP-ribosylation factor-like 3 (ARL3)	NM_004311	1
ATPase, H+ transporting, lysosomal 21kDa, V0 subunit c'' (ATP6V0B)	NM_004047	4
Potassium inwardly-rectifying channel, subfamily J, member 10 (KCNJ10)	NM_002241	1
Adaptor-related protein complex 1, sigma 2 subunit (AP1S2)	NM_003916	1
Calcium channel, voltage-dependent, beta 1 subunit (CACNB1)	NM_000723	1
Potassium voltage-gated channel, KQT-like subfamily, member 4 (KCNQ4)	NM_004700	1
Vitelliform macular dystrophy (Best disease, bestrophin) (VMD2)	NM_004183	151
Other known (Miscellaneous)		
Membrane frizzled-related protein (MFRP)	NM_031433	5
STIP1 homology and U-Box containing protein 1 (STUB1)	NM_005861	1
Hypoxia-inducible factor 1, alpha subunit inhibitor (HIF1AN)	NM_017902	1
Calumenin (CALU)	NM_001219	1
CDC26 subunit of anaphase promoting complex (CDC26)	NM_139286	1
Sterile alpha and HEAT/Armadillo motif protein, ortholog of Drosophila (SARM)	NM_015077	4
Heterogeneous nuclear ribonucleoprotein D-like (HNRPDL)	NM_005463	1
Lysyl oxidase-like 1 (LOXL1)	NM_005576	1
Mitochondrial ribosomal protein S6 (MRPS6)	NM_032476	1
Ubiquitin-like 5 (UBL5)	NM_024292	3
Polymerase (RNA) II (DNA directed) polypeptide E, 25kDa (POLR2E)	NM_002695	1
MutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1)	NM_000249	1
Nucleobindin 1 (NUCB1)	NM_006184	1
Ubiquinol-cytochrome c reductase complex (7.2 kD) (HSPC051)	NM_013387	1
Progesterone receptor membrane component 1 (PGRMC1)	NM_006667	3
Phosphatidylinositol glycan, class S (PIGS)	NM_033198	3
Ribosomal protein S3 (RPS3)	NM_001005	3
Serine/arginine repetitive matrix 2 (SRRM2)	NM_016333	1
Single-stranded DNA binding protein 2 (SSBP2)	NM_012446	1
Protein kinase, lysine deficient 4 (PRKWNK4)	NM_032387	2
Protective protein for beta-galactosidase (galactosialidosis) (PPGB)	NM_000308	1
Serine (or cysteine) proteinase inhibitor, clade A, member 5 (SERPINA5)	NM_000624	3
Staufen, RNA binding protein (Drosophila) (STAU)	NM_017453	1
T-cell leukemia translocation altered gene (TCTA)	NM_022171	1
Unknown		
t-complex-associated-testis-expressed 1-like 1 (TCTEL1)	NM_006519	1
DNA segment, Chr 15, Wayne State University 75, (D15Wsu75e)	XM_039495	3
Aldolase A, fructose-bisphosphate pseudogene 2 (ALDOAP2)	M21191	1
Serine hydrolase-like (SERHL)	NM_170694	3
Tripartite motif-containing 41 (TRIM41)	NM_033549	1
Hypothetical protein (MGC32043)	NM_144582	1
Hypothetical protein (GL009)	NM_032492	1
Mannosidase, beta A, lysosomal-like (MANBAL)	NM_022077	1
Succinate dehydrogenase complex, subunit C, 15kDa (SDHC)	NM_003001	1
Chromosome 6 open reading frame 49 (C6orf49)	NM_017601	2
Chromosome 20 open reading frame 16 (C20orf16)	NM_019025	3
Hypothetical protein (MGC2477)	NM_024099	1
Chloride intracellular channel 6 (CLIC6)	NM_053277	1
Similar to RIKEN cDNA 2610030J16 gene (MGC2541)	NM_080670	1
Hypothetical protein (MGC23937)	NM_145052	1
Hypothetical protein (MGC14161)	NM_032892	1
Chromosome 20 open reading frame 43 (C20orf43)	NM_016407	1
WW domain binding protein 1 (WBP1)	NM_012477	2
Component of oligomeric golgi complex 1 (COG1)	NM_018714	2
Hypothetical protein (MGC10540)	NM_032353	3
Serologically defined breast cancer antigen 84 (SDBCAG84)	NM_015966	4
Seven transmembrane domain protein (NIFIE14)	NM_032635	1
PTD009 protein (PTD009)	NM_016146	1
Similar to DAZ associated protein 2; deleted in azoospermia associated		
Similar to cytochrome b (LOC284125)	XM_210340	1
abhydrolase domain containing 6 (ABHD6)	NM_020676	11
Similar to hypothetical protein PRO2831 (LOC283760)	XM_208826	3
NICE-3 protein (NICE-3)	NM_015449	1

1. (continued)

Gene Name (Gene symbol)	Accession Number	Frequency
C/EBP-induced protein (LOC81558)	NM_030802	1
Mannose-P-dolichol utilization defect 1 (MPDU1)	NM_004870	1
similar to NADH1 (LOC349376)	XM_115915	2
Hypothetical protein (DKFZp761G0122)	NM_152661	15
Six transmembrane epithelial antigen of prostate 2 (STEAP2)	NM_152999	1
Like mouse brain protein E46 (E46L)	NM_013236	2
SPARC related modular calcium binding 2 (SMOC2)	NM_022138	1
ADP-ribosylation factor-like 6 interacting protein (ARL6IP)	NM_015161	1
S-adenosylhomocysteine hydrolase-like 1 (AHCYL1)	NM_006621	2
Basic leucine zipper and W2 domains 2 (BZW2)	NM_014038	1
Similar to RIKEN cDNA 1200015A19 (LOC127262)	XM_072073	1
UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 10 (GALNT10)	NM_017540	1
Hypothetical protein (HSPC111)	NM_016391	1
Hypothetical protein (MGC49942)	BC040148	1
protein 2 (LOC150862)	XM_087029	2
Hypothetical protein BC003515 (LOC154467)	XM_166346	3
Similar to mitochondrial ribosomal protein L55 (LOC128308)	XM_059233	1
Similar to hypothetical protein XP_101622 (LOC161145)	XM_101622	1
LOC119587	XM_058409	11
LOC149464	XM_097645	1
LOC285148	XM_209490	1
LOC286642	XM_212361	1
LOC51256	NM_016495	2
LOC283211	XM_210938	32
LOC91549	XM_039115	1
LOC151361	XM_098048	1
Cystine-knot containing secreted protein (DKFZP564D206)	NM_015464	3
DKFZp434N0419	XM_209007	1
DKFZP547E1010	NM_015607	1
DKFZP564K1964	NM_015544	35
DKFZP564M082	NM_014042	1
KIAA1094	NM_014908	1
KIAA1608	NM_024820	1
KIAA0759	NM_015305	1
KIAA0556	AB011128	2
KIAA0446	AB007915	1
KIAA0445	NM_014675	1
KIAA0157	NM_032182	1
KIAA1189	XM_050508	8
KIAA1576	NM_020927	9
FLJ12076	NM_025187	1
FLJ11305	NM_018386	1
FLJ11756	NM_024606	4
FLJ90119	NM_153347	7
FLJ40773	NM_152666	5
FLJ32069	NM_153033	1
FLJ22055	NM_024779	2
FLJ20580	NM_017887	2
FLJ12089	NM_024552	6
FLJ10803	NM_018224	1
FLJ10504	NM_018116	1
FLJ21032	NM_024906	6
FLJ12287 similar to semaphorins	NM_022367	1

Classification of 341 subtracted ESTs expressed in the human RPE into 18 functional groups. LocusLink accession numbers are shown along with the number of ESTs identified using BlastN program. This supports the view that the RPE is multifunctional.

2. Publication and Presenations

2.1 Thesis-related publication is in preparation

2.2 Presentations at symposium and meeting

2.2.1 Oral Presentation

Faisal M. Fadl Mola, Faisal M. Rahman, Andrea Gehrig, Bernhard H. F. Weber. 2001. Construction of a Relational Database Management System (RDBMS) for the analysis of enriched RPE-derived expressed sequence tags (ESTs). *First International Symposium of the Priority Research Program Age-related Macular Degeneration, Monastery Seeon, Germany*.

2.2.2 Poster Presentation

Faisal M. Fadl Mola, Faisal M. Rahman, Andrea Gehrig, Claudia Keilhauer, Bernhard H. F. Weber. 2002. Construction of a Relational Database Management System (RDBMS) for the analysis of RPE- enriched expressed sequence tags (ESTs). *13th Annual Meeting of the German Society of Human Genetics, Leipzig, Germany*.

3. CURRICULUM VITAE

Personal data

Name: Faisal Mohamed Fadl El Mola

Date of birth: 15.03.1963

Place of birth: Wad Medani, Sudan

Nationality: Sudanese

Marital status: Single

Education and Qualifications

2000 – 2003	PhD student at the Institute of Human Genetics University of Würzburg School of Medicine, Germany
2000	MSc Computer-Based Information Systems School of Computing, Engineering and Technology University of Sunderland, UK
1998-1999	MSc Medical Molecular Biology School of Biosciences, University of Westminster, UK
1997	PG Diploma Computer-Based Information Systems School of Computing, Engineering and Technology University of Sunderland, UK
1983-1988	BSc (Honours) Zoology Faculty of Science, University of Khartoum, Sudan
1980 - 1983	Sudanese High Secondary School Certificate

Work Experience

1993 – 1998	Assistant Information Officer, Islamic Research and Training Institute Jeddah, Saudi Arabia
1990 – 1993	Computer specialist, Saudi Arabian Marketing & Refinery Co. (SAMAREC) Jeddah, Saudi Arabia
1987 – 1990	Practical Teaching Assistant Faculty of Science University of Khartoum, Sudan