Research

Open Access

# Annotation and analysis of 10,000 expressed sequence tags from developing mouse eye and adult retina

Jindan Yu[*], Rafal Farjo[*], Sean P MacNee[*], Wolfgang Baehr[†], Dwight E Stambolian[‡] and Anand Swaroop[*§]

Addresses: [*]Ophthalmology and Visual Science, University of Michigan, 1000 Wall Street, Ann Arbor, MI 48105, USA. [†]Moran Eye Center, University of Utah Health Science Center, Salt Lake City, UT 84132, USA. [‡]Ophthalmology, University of Pennsylvania School of Medicine, Philadelphia, PA 19014, USA. [§]Human Genetics, University of Michigan, 1000 Wall Street, Ann Arbor, MI 48105, USA.

Correspondence: Anand Swaroop. E-mail: swaroop@umich.edu

Published: 22 September 2003

*Genome* ***Biology*** 2003, **4:**R65

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/10/R65

Received: 21 May 2003
Revised: 1 July 2003
Accepted: 19 August 2003

© 2003 Yu *et al.*; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

## Abstract

**Background:** As a biomarker of cellular activities, the transcriptome of a specific tissue or cell type during development and disease is of great biomedical interest. We have generated and analyzed 10,000 expressed sequence tags (ESTs) from three mouse eye tissue cDNA libraries: embryonic day 15.5 (M15E) eye, postnatal day 2 (M2PN) eye and adult retina (MRA).

**Results:** Annotation of 8,633 non-mitochondrial and non-ribosomal high-quality ESTs revealed that 57% of the sequences represent known genes and 43% are unknown or novel ESTs, with M15E having the highest percentage of novel ESTs. Of these, 2,361 ESTs correspond to 747 unique genes and the remaining 6,272 are represented only once. Phototransduction genes are preferentially identified in MRA, whereas transcripts for cell structure and regulatory proteins are highly expressed in the developing eye. Map locations of human orthologs of known genes uncovered a high density of ocular genes on chromosome 17, and identified 277 genes in the critical regions of 37 retinal disease loci. *In silico* expression profiling identified 210 genes and/or ESTs over-expressed in the eye; of these, more than 26 are known to have vital retinal function. Comparisons between libraries provided a list of temporally regulated genes and/or ESTs. A few of these were validated by qRT-PCR analysis.

**Conclusions:** Our studies present a large number of potentially interesting genes for biological investigation, and the annotated EST set provides a useful resource for microarray and functional genomic studies.

## Background

Recent efforts in genomics have accomplished the daunting task of decoding the genome of several species, including human [1,2] and mouse [3]. The current estimate of transcribed genes in the mammalian genome ranges from 35,000 to 45,000, with approximately 99% of mouse genes having homologs in the human genome. Many high-throughput genomics projects have now begun to focus on the identification of cell- and tissue-specific transcriptomes since such gene expression profiles are expected to uncover fundamental

insights into biological processes [4]. To identify genes or cellular pathways that are selectively turned on or off in response to extrinsic factors or intrinsic genetic programs, it is necessary to deduce the catalogue of mRNAs expressed in a specific cell or tissue type at various stages of development, aging and disease.

The vertebrate eye is a key component of the nervous system, the neural retina being responsible for the process of phototransduction - the signal transduction pathway by which light is converted into neural stimuli that result in perceived vision. A systematic evaluation of transcripts and their expression levels at different stages of eye or retinal development should lead to better understanding of underlying regulatory pathways of differentiation and functional maintenance. During the last decade, a number of approaches have been utilized to achieve these tasks. Serial analysis of gene expression (SAGE) provides a catalog of expressed genes of a given tissue through the sequencing of SAGE tags and quantitatively estimates transcript level based on the occurrence of corresponding tags [5-7]. SAGE cataloging has recently been reported for mouse and human retina [8,9]. One caveat of this technique is that tag-to-gene assignments can be ambiguous, since a specific transcript is identified by a short oligonucleotide sequence, usually a 14-20 bp SAGE tag. This is evaded in more traditional expressed sequence tag (EST) generation from cDNA libraries by obtaining a larger tag of 200-600 bp [10,11]. EST generation provides appreciable lengths of sequences of novel genes, which could be deposited into GenBank and complement public databases of ESTs [12-14]. Various computational approaches also rely on EST data to give validity to gene predictions, aid in the detection of functional alternatively spliced transcripts [15,16] and identify tissue-specific genes or candidate disease genes [17-19]. With the availability of microarray technology, slides containing a comprehensive set of ESTs that cover expressed genes of specific tissues during various developmental stages or represent specific cellular pathways, provide a powerful tool for systematic expression profiling, without repetitive sequencing.

A number of ESTs have been isolated from retina and retinal pigment epithelium (RPE) libraries [20-22], from subtracted retina or RPE libraries [23,24] or through computational manipulation and database mining [17,18]. EST generation can be especially valuable in the characterization of uniquely expressed genes or identification of novel candidate genes for retinal disorders [25-27]. Recently, large-scale sequencing has been utilized to gain a clearer picture of the retinal transcriptome through the analyses of an embryonic day 14.5 retinal cDNA library [28] and a set of libraries constructed from different parts of the eye [29]. Corresponding databases of retina/eye ESTs, including RetinaExpress [28], RetBase [19] and the NEIbank database [29], have been generated to serve as useful resources for eye transcriptome consolidation. In addition, microarrays containing small sets of ocular genes/ESTs have been manufactured and used for expression analysis [28,30].

With a goal of obtaining a set of ESTs with deep coverage of expressed genes of ocular tissues, we have generated, annotated and analyzed over 10,000 ESTs derived from three cDNA libraries constructed from developing mouse eye and adult retina. These clones represent a significant resource for producing eye-specific cDNA microarrays (I-GENE microarrays) for comprehensive gene profiling of mouse ocular development and for models of eye disease.

## Results

### Characterization of three mouse cDNA libraries

Three cDNA libraries, namely the M15E, M2PN and MRA libraries, were constructed using total RNAs from mouse embryonic day 15.5 (E15.5) eyes, postnatal day 2.5 (PN2.5) eyes and adult retinas, respectively [30]. No RNA or library amplification or normalization was applied to any of these libraries. A total of 11,057 cDNA clones (4,992 from M15E, 4,128 from M2PN and 1,937 from MRA) were randomly isolated and sequenced to generate ESTs of approximately 500 bp from the 5' end. All ESTs are downloadable from our website [31]. We did not obtain high quality sequences for 1,419 clones: of these, 848 were from the M15E library (Table 1). Further evaluation showed that 577 out of these 848 clones were from the first 24 plates we sequenced from the M15E library, whereas only 134 were from the later 30 plates. Low quality sequences, perhaps caused by sequencing cross-talk, were referred as questionable (group VI, 345 clones), while others, including short sequences, and those with a high-percentage of nucleotide A (adenosine) or N (uncertain read; any of the four nucleotides) and vector sequences, were classified as uninformative (group VII, 1,074 clones). All ESTs of group VI and VII were excluded from further analysis.

A total of 9,638 high quality ESTs (GenBank accession numbers CB839918-CB850489) were compared to the NCBI nr database and the mouse dbEST [32] for homology identification. Based on the analyses, they were categorized into five groups (Table 1). Those with significant homology in NCBI nr databases and representing well-characterized cDNAs were defined as known ESTs (Group I, 4,973 clones). There are 1,956, 2,047 and 970 known ESTs in the M15E, M2PN and MRA libraries, respectively. However, Group I did not include genes of mitochondrial origin or those coding for ribosomal proteins, which were separately classified into group IV and V, respectively. Group II (unknown ESTs, 1,319 clones) refers to ESTs with no match in the NCBI nr database, but with matches in the mouse dbEST. ESTs that have no homology to sequences in both databases were considered as novel ESTs (Group III, 2,341 clones). Approximately 50% of ESTs in each library correspond to known ESTs and 15% to unknown ESTs (Group II). The M15E library has the lowest number in these two categories, reflecting the fact that E15.5 time-point is not

**Table 1**

**Summary of EST analysis**

| | Library | | | | | |
|---|---|---|---|---|---|---|
| | M15E | | M2PN | | MRA | |
| cDNA category | Number of clones | (% ESTs) | Number of clones | (% ESTs) | Number of clones | (% ESTs) |
| High quality ESTs | | | | | | |
| I. Known ESTs | 1,956 | (47.2) | 2,047 | (55.3) | 970 | (54.1) |
| II. Unknown ESTs | 463 | (11.2) | 557 | (15.0) | 299 | (16.7) |
| III. Novel ESTs | 1,200 | (29.0) | 792 | (21.4) | 349 | (19.5) |
| IV. Mitochondrial DNA | 248 | (6.0) | 202 | (5.5) | 132 | (7.4) |
| V. Ribosomal RNA, protein | 277 | (6.7) | 103 | (2.8) | 43 | (2.4) |
| Subtotal | 4,144 | (100) | 3,701 | (100) | 1,793 | (100) |
| VI. Questionable sequences * | 137 | | 143 | | 65 | |
| VII. Uninformative sequences * | 711 | | 284 | | 79 | |
| Total | 4,992 | | 4,128 | | 1,937 | |

* Sequences in these categories were not deposited in the GenBank database and were excluded from further analysis.

as well studied. The M15E library also contains the largest fraction of novel ESTs (29%), relative to 21.4% in M2PN and 19.5% in MRA, demonstrating the under-representation of E15.5 eye ESTs in public databases.

### Gene annotation

To gain a better understanding of our EST sets, we performed a series of analyses (Figure 1). ESTs were masked to eliminate repeat sequences and examined for high quality. A number of sequence similarity searches were executed to compare every EST to those in public or in our local databases. For ESTs with known-gene matches in public databases, functional annotation was retrieved from NCBI UniGene [33] and LocusLink [34]. Of the 4,973 known ESTs, 65% show corresponding UniGene and LocusLink entries, and 48% have matching human orthologs and identified chromosomal locations.

To assist in the analyses of tissue expression patterns of the 1,319 unknown ESTs, we retrieved from the NCBI Entrez nucleotide database [32] all dbEST entries with significant homology to our clones and determined their original tissue of isolation. Cross-comparison of every EST to the complete set of ESTs in our local database provided information regarding their redundancy and level of expression in each of the three libraries. Gene annotation, along with sequence quality assessment, was recorded in a local database with hyperlinks to NCBI Entrez nucleotide, UniGene and LocusLink databases (Figure 2). A complete list of all clones and corresponding annotations is available on our website [31].

### EST clustering

To estimate the number of unique transcripts represented by our ESTs, known, unknown or novel ESTs were clustered within each library. We did not consider redundancy between libraries, as our purpose was to assess the amount of redundant sequences within an individual library and to compare EST frequency between libraries. ESTs matching known genes were placed in groups based on common nucleotide hits in the NCBI nr database. The results showed 3,604 unique genes (1,382 in M15E, 1,443 in M2PN and 779 in MRA) representing these 4,973 ESTs. Less than 50 clusters contained more than five ESTs and a majority (3,016) consisted of a single EST (Figure 3a). The largest cluster in the M15E, M2PN and MRA libraries contained, respectively, 45, 50 and 37 ESTs corresponding to the same gene. Table 2 includes highly-expressed genes of each library. Crystallin genes were excluded from this list since the inclusion of lens in the sample RNA used for library construction could bias their abundance in the M15E and M2PN libraries. In accordance with previous observations [28], translation factors, cell structure/cytoskeletal proteins and housekeeping genes are among the most abundant, especially in the M15E and M2PN libraries. In the MRA library, phototransduction genes are among those highly expressed, consistent with the primary functional responsibility of the mature retina. Of the 12 most abundant genes, ten are known to play important roles in retinal function; these include rhodopsin (37 occurrences), alpha-transducin 1 (*Gnat1*) (nine occurrences), arrestin (*Sag*) (six occurrences), glutamine synthetase (five occurrences), unc119 homolog (*Unc119 h*, *Hrg4*) (five occurrences), interphotoreceptor retinoid-binding protein (five occurrences) and four occurrences each of tubby like protein 1 (*Tulp1*), phosducin, guanylate cyclase activator 1a (*Guca1a*) and peripherin 2 (*Rds*). We conclude that genes/ESTs from these three cDNA libraries are representative of their source tissues and their redundancy could be utilized to construct *in silico* expression profiles.

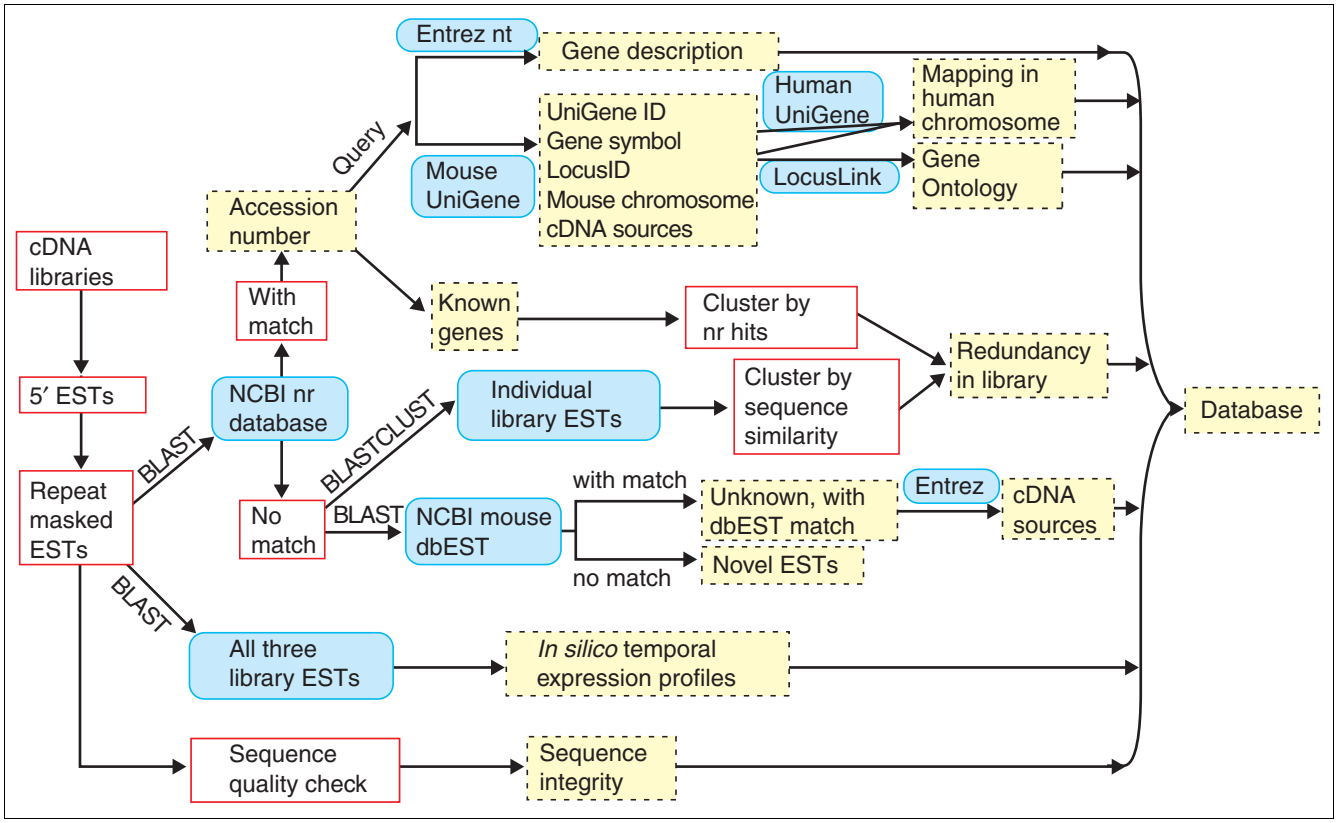


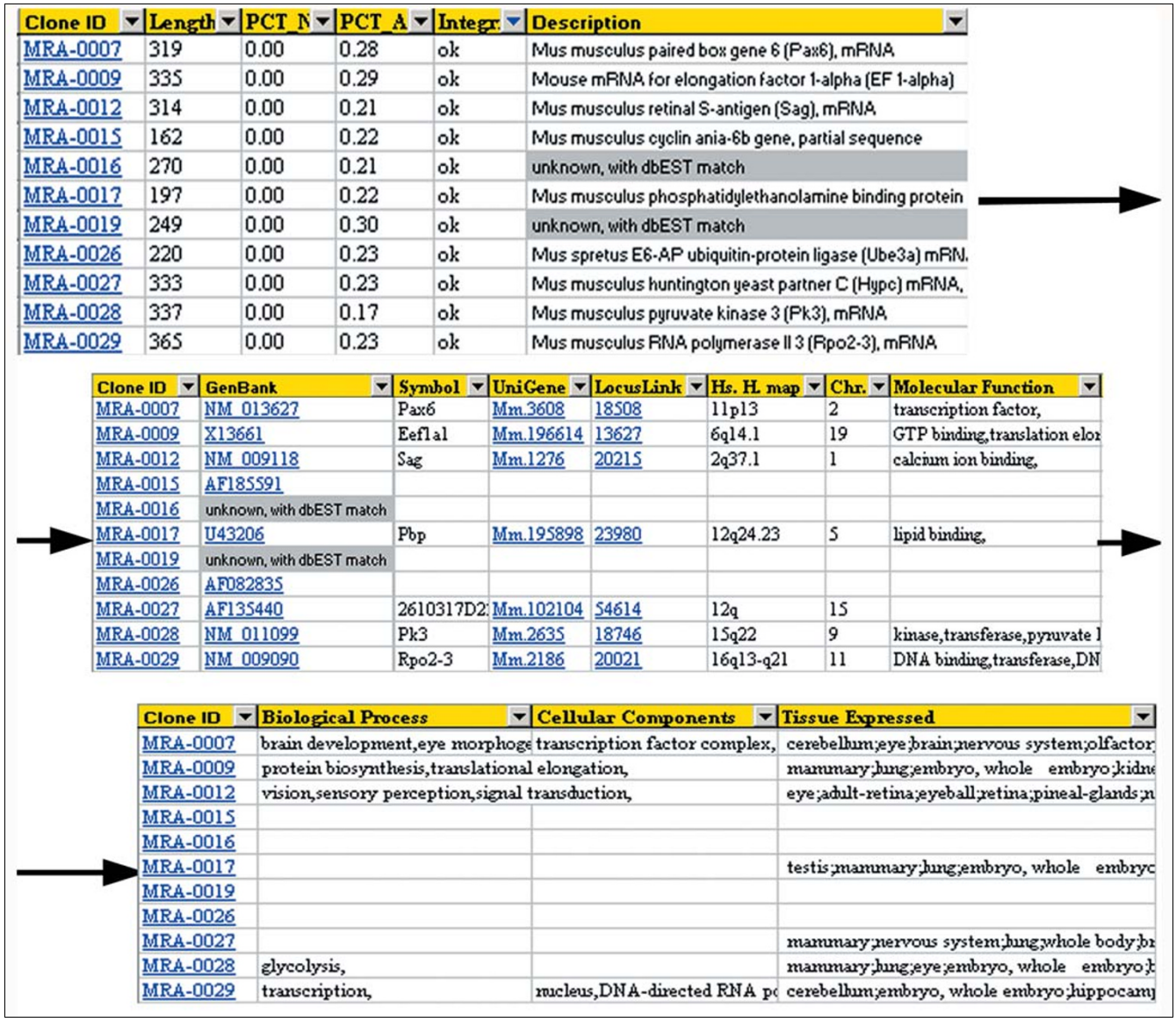

**Figure 1**
Schematic representation of the EST analysis and annotation process. ESTs from the 5' end of cDNAs were repeat masked and checked for sequence integrity. They were BLAST-searched against the NCBI nr database, the mouse dbEST and our local database of EST sets from each library (individual library ESTs) or the entire collection of 9,638 ESTs (all three library ESTs). High-level functional annotation was achieved by searching the NCBI databases, including Entrez nt, LocusLink, mouse and human UniGene database. Databases that were BLASTed or queried are indicated by solid blue, rounded rectangles. Annotated data (highlighted by dashed, yellow rectangles) for every EST can be accessed at [31].

Unknown and novel ESTs were clustered based on sequence similarity. A total of 159 clusters, composed of 404 ESTs, were generated, while the remaining 3,256 ESTs did not cluster, resulting in a total of 3,415 (93%) non-redundant clusters for the total 3,660 ESTs (Figure 3b). Only five ESTs from all three libraries have redundancy that is higher than three. Therefore, the number of unique transcripts in our clone set, including all known, unknown and novel ESTs, was estimated to be up to 7,019 (81% of all clustered genes/ESTs), with 2,909 from M15E, 2,705 from M2PN and 1,405 from the MRA library. This estimation excludes mitochondrial, ribosomal and low-quality sequences. The redundancy of known ESTs in the libraries is relatively higher than that of unknown and novel ESTs, which may reflect easier identification of abundantly expressed genes, near completion of genomic sequencing in human and mouse, and recent focus on functional characterization.

### Functional distribution of known ESTs

Of the 4,973 ESTs matching to known genes in the nr database, 1,142 correspond to cDNAs with unspecified function and 482 have homology with genomic sequences. The remaining 3,349 known ESTs can be divided into ten groups based on their putative functions (Figure 4a). The largest two groups include proteins involved in cell signaling/communication (17%) and those involved in protein expression/regulation/processing (15%). Following these were functional groups including crystallin-family genes (14%), and those involved in cell structure/motility/extracellular matrix (13%) and gene regulation/transcription factors (12%).

A detailed breakdown of functional groups by cDNA library (Figure 4b,c,d) demonstrated that phototransduction-specific genes are significantly more abundant in the MRA library, in concordance with functional activity at this stage. We also observed that crystallin genes are highly represented in the M15E and M2PN libraries, which is due to the fact that these two libraries were constructed from whole eye, while MRA was constructed from retina only. In general, the fraction of genes devoted to cellular functions in other categories did not deviate significantly from the overall pattern. At developmental stages, more genes appear to be involved in cell structure/motility/extracellular matrix (14%) than in mature retina (9%). There is also a slight decrease in genes for



| Clone ID | Length | PCT_N | PCT_A | Integr | Description |
|---|---|---|---|---|---|
| MRA-0007 | 319 | 0.00 | 0.28 | ok | Mus musculus paired box gene 6 (Pax6), mRNA |
| MRA-0009 | 335 | 0.00 | 0.29 | ok | Mouse mRNA for elongation factor 1-alpha (EF 1-alpha) |
| MRA-0012 | 314 | 0.00 | 0.21 | ok | Mus musculus retinal S-antigen (Sag), mRNA |
| MRA-0015 | 162 | 0.00 | 0.22 | ok | Mus musculus cyclin ania-6b gene, partial sequence |
| MRA-0016 | 270 | 0.00 | 0.21 | ok | unknown, with dbEST match |
| MRA-0017 | 197 | 0.00 | 0.22 | ok | Mus musculus phosphatidylethanolamine binding protein |
| MRA-0019 | 249 | 0.00 | 0.30 | ok | unknown, with dbEST match |
| MRA-0026 | 220 | 0.00 | 0.23 | ok | Mus spretus E6-AP ubiquitin-protein ligase (Ube3a) mRN. |
| MRA-0027 | 333 | 0.00 | 0.23 | ok | Mus musculus huntington yeast partner C (Hypc) mRNA, |
| MRA-0028 | 337 | 0.00 | 0.17 | ok | Mus musculus pyruvate kinase 3 (Pk3), mRNA |
| MRA-0029 | 365 | 0.00 | 0.23 | ok | Mus musculus RNA polymerase II 3 (Rpo2-3), mRNA |

| Clone ID | GenBank | Symbol | UniGene | LocusLink | Hs. H. map | Chr. | Molecular Function |
|---|---|---|---|---|---|---|---|
| MRA-0007 | NM_013627 | Pax6 | Mm.3608 | 18508 | 11p13 | 2 | transcription factor, |
| MRA-0009 | X13661 | Eef1a1 | Mm.196614 | 13627 | 6q14.1 | 19 | GTP binding,translation elo |
| MRA-0012 | NM_009118 | Sag | Mm.1276 | 20215 | 2q37.1 | 1 | calcium ion binding, |
| MRA-0015 | AF185591 | | | | | | |
| MRA-0016 | unknown, with dbEST match | | | | | | |
| MRA-0017 | U43206 | Pbp | Mm.195898 | 23980 | 12q24.23 | 5 | lipid binding, |
| MRA-0019 | unknown, with dbEST match | | | | | | |
| MRA-0026 | AF082835 | | | | | | |
| MRA-0027 | AF135440 | 2610317D2 | Mm.102104 | 54614 | 12q | 15 | |
| MRA-0028 | NM_011099 | Pk3 | Mm.2635 | 18746 | 15q22 | 9 | kinase,transferase,pyruvate |
| MRA-0029 | NM_009090 | Rpo2-3 | Mm.2186 | 20021 | 16q13-q21 | 11 | DNA binding,transferase,DN |

| Clone ID | Biological Process | Cellular Components | Tissue Expressed |
|---|---|---|---|
| MRA-0007 | brain development,eye morphoge | transcription factor complex, | cerebellum;eye;brain;nervous system;olfactor |
| MRA-0009 | protein biosynthesis,translational elongation, | | mammary;lung;embryo, whole embryo;kidn |
| MRA-0012 | vision,sensory perception,signal transduction, | | eye;adult-retina;eyeball;retina;pineal-glands;n |
| MRA-0015 | | | |
| MRA-0016 | | | |
| MRA-0017 | | | testis;mammary;lung;embryo, whole embryc |
| MRA-0019 | | | |
| MRA-0026 | | | |
| MRA-0027 | | | mammary;nervous system;lung;whole body;br |
| MRA-0028 | glycolysis, | | mammary;lung;eye;embryo, whole embryo;h |
| MRA-0029 | transcription, | nucleus,DNA-directed RNA p | cerebellum;embryo, whole embryo;hippocamp |

**Figure 2**
Screen shot of I-GENE database of annotated ESTs. Clones are hyperlinked to their sequences. GenBank, UniGene and LocusLink IDs are hyperlinked to the corresponding web page. The integrity of each sequence was assessed by sequence length, percentage of Ns (PCT_N) and percentage of As (PCT_A) contained. Arrows are used to indicate continuation of the table for each clone ID.

gene regulation/transcription and for protein expression/regulation/processing in the adult retina. In contrast, mature retina has more genes involved in metabolism (10%) compared to 6-7% in developmental libraries. Overall, the analysis of functional distribution of known ESTs in the three libraries is consistent with the established developmental and functional role at specific stages.

### Candidate ocular disease genes

To identify new candidate genes for ocular, in particular retinal, diseases, we have analyzed the chromosomal locations of all ESTs with known human gene matches. We first obtained the chromosomal location for each mouse gene, followed by the determination of the location of corresponding human ortholog if available. Of the 3,604 unique known genes, we were able to obtain mouse chromosomal locations for 1,964 genes, 1,522 of which have human orthologs and mapping information based on the UniGene database. Distribution by chromosome of these genes in mouse and human genome is given in Table 3. Expected genes per human chromosome were computed based on the Human Gene Map database [35]. Chi-square analysis indicated that the observed frequency of ocular genes by chromosome deviates significantly from the expectation for all three libraries ($\chi^2$ = 72.2, 31.9, 67.46 for the M15E, M2PN and MRA libraries, respectively). The ratio of observed versus expected frequency by

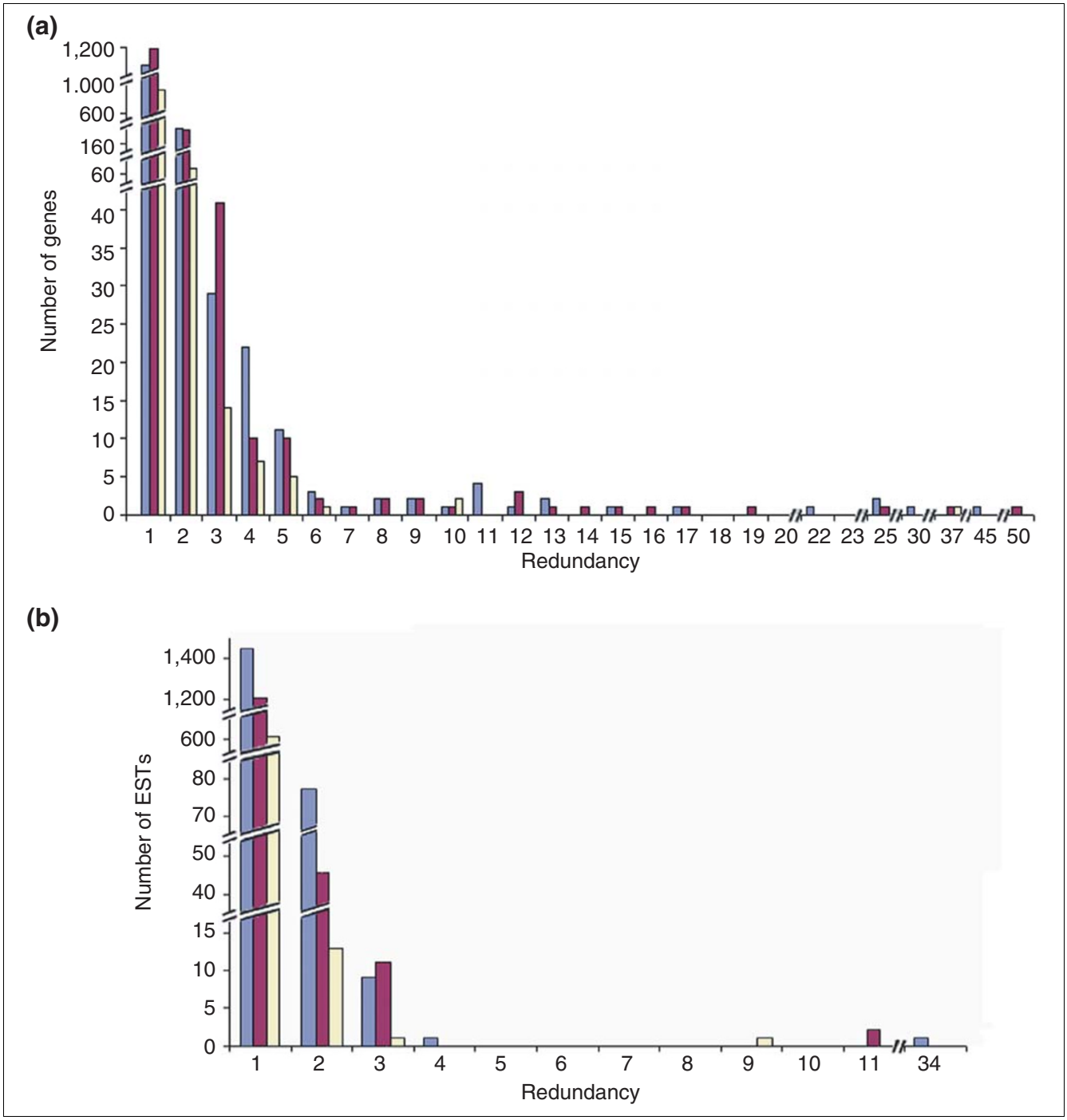



**Figure 3**
Histograms showing the number of genes or ESTs at each level of redundancy in the three libraries. Redundancy of **(a)** known genes and **(b)** unknown or novel ESTs are shown for the M15E (blue), M2PN (purple) and MRA (yellow) libraries.

chromosomes revealed that the ocular gene density on chromosome 17 is significantly higher than expected ($p < 0.002$), but there appears to be no clustering in any specific chromosomal region. The X-chromosome has a marginally higher density of ocular genes from the M2PN and MRA libraries, whereas chromosomes 4 and 10 have marginally fewer ocular genes.

The RetNet database [36] was searched to determine whether any of these genes might serve as candidates for mapped but

**Table 2**

**Highly-expressed genes in the M15E, M2PN and MRA libraries, excluding crystallin genes**

| Accession number | Gene name | Number of occurrences | %Total |
|---|---|---|---|
| M15E library | | | |
| NM_008218 | Hemoglobin alpha, adult chain 1 (*Hba-a1*) | 45 | 2.30 |
| M22432 | Protein synthesis elongation factor Tu (*eEF-Tu*, *eEf-1-alpha*) | 30 | 1.53 |
| NM_008220 | Hemoglobin, beta adult major chain (*Hbb-b1*) | 22 | 1.12 |
| NM_011653 | Tubulin alpha 1 (*Tuba1*) | 13 | 0.66 |
| NM_008084 | Glyceraldehyde-3-phosphate dehydrogenase (*Gapd*) | 9 | 0.46 |
| NM_008302 | Heat shock protein, 84 kDa 1 (*Hsp84-1*) | 8 | 0.41 |
| NM_008972 | Prothymosin alpha (*Ptma*) | 6 | 0.31 |
| NM_025586 | RIKEN cDNA 2510008H07 gene (2510008H07Rik) | 6 | 0.31 |
| NM_026055 | RIKEN cDNA 2810465O16 gene (2810465O16Rik) | 6 | 0.31 |
| K01173 | Dog (canine) chymotrypsin | 5 | 0.26 |
| NM_007393 | Actin, beta, cytoplasmic (*Actb*) | 5 | 0.26 |
| NM_011664 | Ubiquitin B (*Ubb*) | 5 | 0.26 |
| NM_023119 | Enolase 1, alpha non-neuron (*Eno1*) | 5 | 0.26 |
| NM_023123 | H19 fetal liver mRNA (*H19*) | 5 | 0.26 |
| NM_025379 | cytochrome c oxidase subunit VIIb (*Cox7b*) | 5 | 0.26 |
| NM_053075 | RAS-homolog enriched in brain (*Rheb*) | 5 | 0.26 |
| M2PN library | | | |
| M22432 | Protein synthesis elongation factor Tu (*eEF-Tu*, *eEf-1-alpha*) | 12 | 0.59 |
| NM_011664 | Ubiquitin B (*Ubb*) | 10 | 0.49 |
| NM_010240 | Ferritin light chain 1 (*Ftl1*) | 9 | 0.44 |
| NM_009751 | Beaded filament structural protein in lens-CP94 (*Bfsp1*) | 8 | 0.39 |
| NM_009242 | Secreted acidic cysteine rich glycoprotein (*Sparc*) | 6 | 0.29 |
| NM_021278 | Thymosin, beta 4, X chromosome (*Tmsb4x*) | 6 | 0.29 |
| AK013100 | 10, 11 days embryo whole body cDNA, RIKEN clone:2810417D04 | 5 | 0.24 |
| BE657894 | GM700004A10F4 Gm-r1070 Glycine max cDNA clone Gm-r1070-1399 | 5 | 0.24 |
| NM_007393 | Melanoma X-actin (*Actx*) | 5 | 0.24 |
| NM_007687 | Cofilin 1, non-muscle (*Cfl1*) | 5 | 0.24 |
| NM_008084 | Glyceraldehyde-3-phosphate dehydrogenase (*Gapd*) | 5 | 0.24 |
| NM_018785 | Formin binding protein 3 (*Fnbp3*) | 5 | 0.24 |
| MRA library | | | |
| BC013125 | Rhodopsin (*Rho*) | 37 | 3.81 |
| NM_008084 | Glyceraldehyde-3-phosphate dehydrogenase (*Gapd*) | 10 | 1.03 |
| NM_008140 | Guanine nucleotide binding protein, alpha transducing 1 (*Gnat1*) | 10 | 1.03 |
| NM_009118 | Retinal S-antigen (*Sag*) | 6 | 0.62 |
| NM_008131 | Glutamine synthetase (*Glns*) | 5 | 0.52 |
| NM_010480 | Heat shock protein, 86 kDa 1 (*Hsp86-1*) | 5 | 0.52 |
| NM_011676 | Unc119 homolog (*C. elegans*) (*Unc119 h*) | 5 | 0.52 |
| X69523 | *R. rattus* mRNA for interphotoreceptor retinoid-binding protein (*Rbp3*) | 5 | 0.52 |
| AF105711 | Tubby like protein 1 (*Tulp1*) | 4 | 0.41 |
| L08075 | Domesticus phosducin | 4 | 0.41 |
| NM_008189 | *Mus musculus* guanylate cyclase activator 1a (retina) (*Guca1a*) | 4 | 0.41 |
| NM_008938 | *Mus musculus* peripherin 2 (*Prph2*) | 4 | 0.41 |

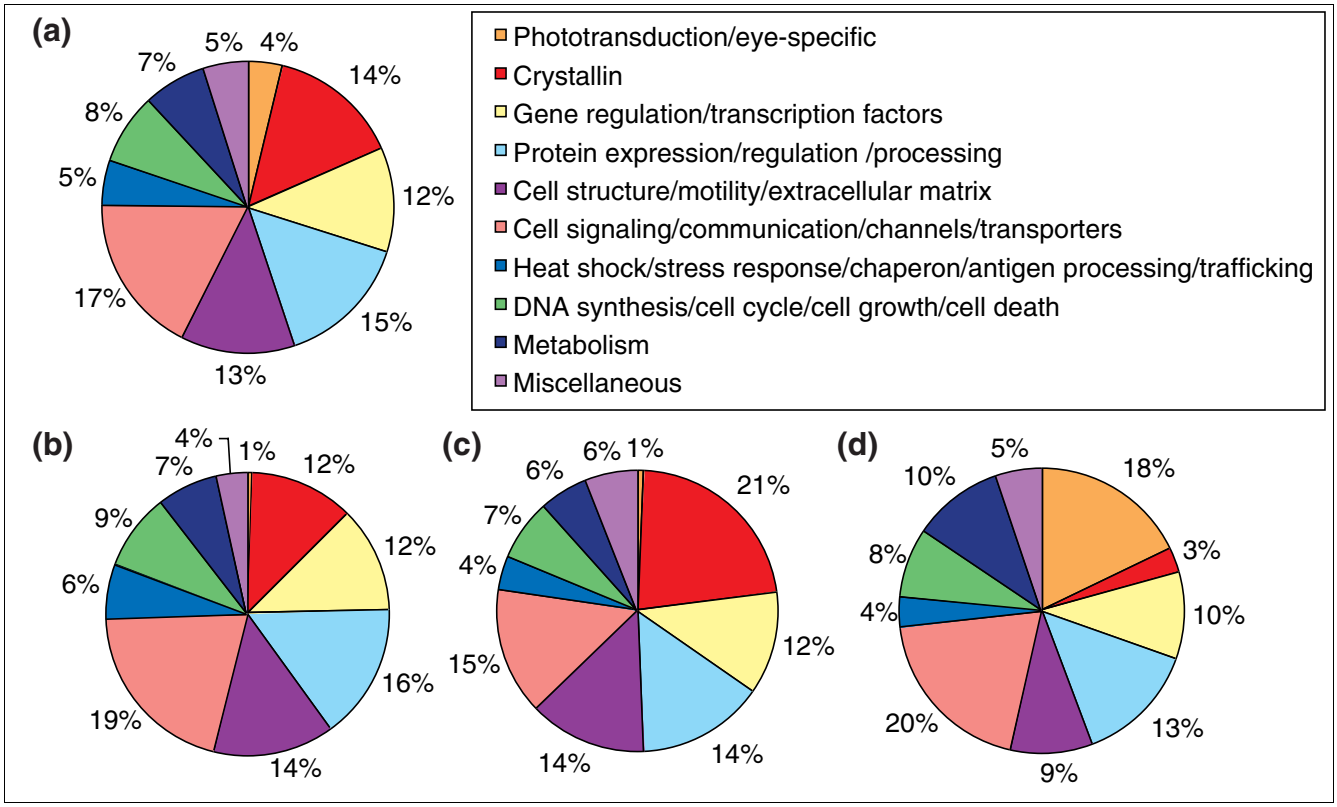



**Figure 4**
Functional categorization of ESTs with known-gene match. **(a)** Categorization for the total 3,349 known ESTs across the three libraries, **(b)** the breakdown for M15E, **(c)** the breakdown for M2PN and **(d)** the breakdown for the MRA library. The number next to each category indicates the percentage of that particular group in the total ten classes.

as yet uncloned retinal disease loci. Of the 2,358 human ortholog ESTs, 641 ESTs, representing 277 non-redundant genes, were localized to one of the mapped retinal disease regions. A complete list of these genes is available in Additional data file 1. Table 4 summarizes the number of genes mapping into each retinal disease locus, with one example gene listed. ESTs are localized to the cytogenetic locations of a total of 37 retinal disease loci. The RP26 region includes 142 ESTs, but this may be due to the fact that several crystallin genes with high abundance in our libraries are localized in this region.

### *In silico* expression profiling

Since the cDNA clones were randomly selected for sequencing, their relative abundance in an unamplified library may represent the corresponding transcript levels.

#### *Temporal expression profiles*

To determine expression level of ESTs in the M15E, M2PN and MRA libraries, each sequence was queried against our entire collection of 9,638 sequences using a local BLAST database. For every query, the number of matching ESTs from the three libraries was recorded. Relative frequencies of ESTs were computed by totaling the number of matching genes/ESTs within this library (number of occurrences) and dividing it by the total number of ESTs sequenced (excluding low-quality sequences) for this library. *In silico* temporal expression profiles were then constructed for each EST based on their relative frequencies in the three libraries. In Table 5, we selectively show a number of unknown ESTs with age-restricted patterns of expression. As expected, phototransduction genes are highly represented in the MRA library (data not shown). Rhodopsin and its homologs represent 1.3% of MRA clones, but are absent in the other two libraries. We observed a number of genes/ESTs that are highly expressed in both the M2PN and MRA libraries and are restricted to these postnatal stages. These include translation repressor NAT1 (0% in M15E, 2.0% in M2PN and 1.4% in MRA), hairy and enhancer of split 6 (0% in M15E, 2.0% in M2PN and 1.5% in MRA; data not shown), and a number of unknown ESTs (Table 5). These genes may be specifically relevant to postnatal ocular/retinal functions. Many members of crystallin genes are highly expressed in the M2PN library. Considering both M15E and M2PN were constructed from eye tissues, the greater percentage of these genes in M2PN may reflect the development of lens at PN2. Ribosomal proteins are found to be more abundant in the M15E library, consistent with previous observations in the E14.5 library [28]. Examples of ESTs

**Table 3**

**Ocular gene distribution in mouse and human chromosomes**

| | | | | Human ortholog of mouse ocular genes | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mouse genes | | | Observed | | | Expected | | | Chi-square | | |
| Chromosome | M15E | M2PN | MRA | M15E | M2PN | MRA | M15E | M2PN | MRA | M15E | M2PN | MRA |
| 1 | 47 | 49 | 29 | 53 | 52 | 26 | 64.2 | 60.7 | 32.7 | 1.95 | 1.24 | 1.38 |
| 2 | 61 | 71 | 29 | 51 | 51 | 19 | 46.5 | 44.0 | 23.7 | 0.43 | 1.12 | 0.94 |
| 3 | 44 | 39 | 23 | 36 | 31 | 16 | 41.5 | 39.3 | 21.2 | 0.74 | 1.74 | 1.26 |
| 4 | 61 | 36 | 21 | 20 | 19 | 12 | 30.5 | 28.8 | 15.5 | 3.60 | 3.33 | 0.80 |
| 5 | 46 | 35 | 21 | 27 | 36 | 7 | 31.5 | 29.8 | 16.1 | 0.65 | 1.30 | 5.11 |
| 6 | 36 | 38 | 19 | 23 | 28 | 33 | 39.0 | 36.9 | 19.9 | 6.58 | 2.14 | 8.64 |
| 7 | 66 | 61 | 26 | 27 | 28 | 6 | 32.9 | 31.1 | 16.8 | 1.05 | 0.30 | 6.90 |
| 8 | 40 | 36 | 21 | 20 | 25 | 15 | 24.9 | 23.5 | 12.7 | 0.95 | 0.10 | 0.43 |
| 9 | 39 | 48 | 23 | 25 | 19 | 10 | 25.7 | 24.3 | 13.1 | 0.02 | 1.16 | 0.74 |
| 10 | 39 | 41 | 30 | 19 | 19 | 4 | 28.3 | 26.7 | 14.4 | 3.04 | 2.23 | 7.52 |
| 11 | 85 | 62 | 42 | 44 | 35 | 19 | 36.2 | 34.2 | 18.4 | 1.69 | 0.02 | 0.02 |
| 12 | 24 | 16 | 14 | 33 | 35 | 21 | 32.7 | 30.9 | 16.7 | 0.00 | 0.55 | 1.14 |
| 13 | 21 | 29 | 12 | 14 | 10 | 5 | 14.5 | 13.7 | 7.4 | 0.02 | 1.00 | 0.77 |
| 14 | 24 | 33 | 16 | 22 | 22 | 18 | 21.6 | 20.4 | 11.0 | 0.01 | 0.13 | 4.46 |
| 15 | 28 | 28 | 23 | 24 | 23 | 15 | 21.2 | 20.0 | 10.8 | 0.37 | 0.44 | 1.63 |
| 16 | 20 | 16 | 14 | 31 | 16 | 7 | 17.5 | 16.5 | 8.9 | **10.40** | 0.02 | 0.41 |
| 17 | 41 | 31 | 25 | 48 | 34 | 25 | 26.0 | 24.6 | 13.3 | **18.51** | 3.58 | **10.36** |
| 18 | 16 | 14 | 3 | 3 | 7 | 2 | 10.8 | 10.2 | 5.5 | 5.61 | 1.00 | 2.22 |
| 19 | 33 | 26 | 18 | 25 | 28 | 18 | 23.0 | 21.7 | 11.7 | 0.18 | 1.82 | 3.38 |
| 20 | NA | NA | NA | 18 | 19 | 7 | 15.6 | 14.8 | 8.0 | 0.36 | 1.22 | 0.12 |
| 21 | NA | NA | NA | 13 | 7 | 5 | 6.3 | 5.9 | 3.2 | 7.14 | 0.19 | 1.00 |
| 22 | NA | NA | NA | 21 | 15 | 9 | 11.6 | 11.0 | 5.9 | 7.52 | 1.45 | 1.58 |
| X | 27 | 30 | 18 | 23 | 27 | 17 | 18.0 | 17.0 | 9.2 | 1.38 | 5.84 | 6.65 |
| Total | 798 | 739 | 427 | 620 | 586 | 316 | 620 | 586 | 316 | **72.20** | 31.90 | **67.46** |

Chi-square tests with 1 degree of freedom ($df = 1$) for each chromosome and with $df = 22$ for the totals were applied to compare the observed distribution of ocular genes with expected distribution calculated based on Human Gene Map. Significant Chi-square values with $p < 0.002$ are indicated in bold. Marginally significant Chi-square values with $p < 0.1$ are underscored, if genes from all three libraries showed the same trend of higher or lower density.

preferentially expressed at specific time-points are listed in Table 5, while *in silico* temporal expression profiles of all ESTs are available at our website [31].

*Tissue expression profiles*

To identify potential retina-enriched genes, cDNA sources of known ESTs and tissue origins of homologs of unknown ESTs were examined. Of 3,189 known ESTs (1,290 from M15E, 1,259 from M2PN and 638 from MRA) with cDNA sources annotated, we have identified 110 unique genes that are expressed preferentially in eye, retina, pineal and brain tissues, including 28 from M15E, 30 from M2PN and 52 from MRA. Roughly half (26) of the retina-enriched genes from the MRA library are known to be involved in retinal function. Table 6 lists these potential retina-enriched genes. This table excludes known phototransduction genes and crystallins.

In the case of unknown ESTs, those matching predominantly to ESTs from ocular tissues were considered as eye-enriched (Table 7). We examined the feasibility of this approach using sequences for rhodopsin, *unc119h* and *Rlbp1* (*Cralbp*) (cellular retinaldehyde-binding protein 1). The rhodopsin sequence has 250 homology ESTs in the NCBI mouse dbEST database, all of which were isolated from retina-related tissues. We also observed that 34 out of 68 of *unc119h* and 19 out of 23 of *Rlbp1* homology ESTs were identified in ocular tissues. A total of 100 unknown ESTs from our study were found to be ocular specific, as over 70% of their homologous ESTs in the NCBI

**Table 4**

**Candidate ocular disease genes identified in three mouse retina/eye cDNA libraries**

| Disease | Location | Number of ESTs in interval | Number of unique genes | Example accession number | Gene name |
|---|---|---|---|---|---|
| AA | 11p15 | 60 | 19 | NM_015814 | Dickkopf homolog 3 (*Xenopus laevis*) |
| ARMD1 | 1q25-q31 | 17 | 8 | BC023001 | Regulator of G-protein signaling 2 |
| AXPC1 | 1q31-q32 | 18 | 9 | NM_008131 | Glutamine synthetase (*Glns*) |
| BBS5 | 2q31 | 8 | 4 | AK019448 | Procollagen, type III, alpha 1 |
| BCD | 4q35-qter | 2 | 1 | U27315 | Adenine nucleotide translocase-1 (*Ant1*) |
| CACD | 17p13 | 26 | 16 | BC008093 | Eukaryotic translation initiation factor 5A |
| COD2, XLPCD | Xq27 | 2 | 1 | NM_008031 | Fragile X mental retardation syndrome 1 homolog (*Fmr1*) |
| CORD1 | 18q21.1-q21.3 | 2 | 1 | NM_009190 | Vacuolar protein sorting 4b (yeast) |
| CORD7 | 6q | 83 | 29 | NM_007865 | Delta-like 1 (*Drosophila*) |
| CORD8 | 1q12-q24 | 30 | 20 | NM_011342 | SEC22 vesicle trafficking protein-like 1 (*S. cerevisiae*) |
| CORD9 | 8p11 | 5 | 3 | NM_031158 | Ankyrin 1, erythroid |
| CYMD | 7p21-p15 | 20 | 10 | AK002910 | Chromobox homolog 3 (*Drosophila* HP1 gamma) |
| EVR3 | 11p13-p12 | 9 | 6 | NM_013627 | Paired box gene 6 (*Pax6*) |
| LCA3 | 14q24 | 11 | 8 | NM_007701 | *C. elegans* ceh-10 homeo domain containing homolog |
| LCA5 | 6q11-q16 | 45 | 6 | NM_009945 | Cytochrome c oxidase, subunit VIIa 3 (*Cox7a3*) |
| MCDR1 | 6q14-q16.2 | 41 | 4 | AK019500 | NS1-associated protein 1 |
| MRST | 15q24 | 12 | 6 | NM_024431 | Testis expressed gene 189 |
| OPA2 | Xp11.4-p11.2 | 26 | 10 | NM_009457 | Ubiquitin-activating enzyme E1, Chr X |
| PRD | Xp11.3-p11.23 | 19 | 8 | NM_013680 | Synapsin I |
| RCD1 | 6q25-q26 | 8 | 3 | AF147785 | Zinc finger protein ZAC1 (*Zac1*) |
| RNANC | 10q21 | 3 | 2 | X16461 | Cell division cycle 2 homolog A (*S. pombe*) |
| RP17 | 17q22 | 10 | 4 | NM_009296 | Suppressor of Ty 4 homolog (*S. cerevisiae*) |
| RP22 | 16p12.1-p12.3 | 6 | 4 | NM_007672 | Cerebellar degeneration-related 2 |
| RP24 | Xq26-q27 | 4 | 2 | NM_016697 | Glypican 3 |
| RP25 | 6cen-q15 | 45 | 6 | NM_009945 | Cytochrome c oxidase, subunit VIIa 2 |
| RP26 | 2q31-q33 | 142 | 20 | NM_022988 | Ngg1 interacting factor 3-like 1 (*S. pombe*) |
| RP28 | 2p11-p16 | 26 | 17 | NM_009837 | Chaperonin subunit 4 (delta) (*Cct4*) |
| RP29 | 4q32-q34 | 4 | 3 | NM_025436 | Sterol-C4-methyl oxidase-like (*Sc4mol*) |
| STGD4 | 4p | 9 | 7 | NM_013457 | Adducin 1 (alpha) |
| USH1A, USH1 | 14q32 | 13 | 6 | NM_020494 | DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 13 (RNA helicase A) |
| USH1E | 21q21 | 2 | 1 | BC013562 | Similar to E4tf1-60 transcription factor |
| USH1G | 17q24-q25 | 39 | 26 | AK012619 | SMT3 (supressor of mif two, 3) homolog 2 (*S. cerevisiae*) |
| USH2B | 3p24.2-p23 | 4 | 3 | NM_009455 | Ubiquitin-conjugating enzyme E2E 1, UBC4/5 homolog (yeast) |
| USH2C | 5q14-q21 | 7 | 5 | NM_010151 | Nuclear receptor subfamily 2, group F, member 1 |
| VRNI | 11q13 | 45 | 24 | NM_023131 | Ras and a-factor-converting enzyme 1 homolog (*S. cerevisiae*) (*Rce1*) |
| WFS2 | 4q22-q24 | 13 | 9 | NM_007917 | Eukaryotic translation initiation factor 4E (*Eif4e*) |
| WGN1, ERVR | 5q13-q14 | 13 | 9 | AY037837 | Single-stranded DNA binding protein 2 |

nucleotide database were originally isolated from ocular tissue. A complete list of tissue expression patterns of unknown ESTs is available at our website [31].

### qRT-PCR expression profiling of selected clones

To validate *in silico* expression profiles for different developmental stages and tissues, we performed qRT-PCR analysis

**Table 5**

***In silico* temporal expression profiles of selected ESTs**

| Clone ID | Accession number | Number in M15E | Number in M2PN | Number in MRA | % in M15E | % in M2PN | % in MRA |
|---|---|---|---|---|---|---|---|
| M2PN-0339 | CB844781 | 0 | 74 | 26 | 0 | 2.00 | 1.45 |
| M2PN-0448 | CB844886 | 1 | 74 | 25 | 0.02 | 2.00 | 1.39 |
| M2PN-0649 | CB845079 | 0 | 74 | 26 | 0 | 2.00 | 1.45 |
| M2PN-0722 | CB845148 | 0 | 74 | 26 | 0 | 2.00 | 1.45 |
| M2PN-0316 | CB844758 | 0 | 73 | 27 | 0 | 1.97 | 1.51 |
| M2PN-0376 | CB844818 | 0 | 73 | 27 | 0 | 1.97 | 1.51 |
| M2PN-0391 | CB844833 | 0 | 73 | 27 | 0 | 1.97 | 1.51 |
| M2PN-0438 | CB844876 | 0 | 73 | 27 | 0 | 1.97 | 1.51 |
| M2PN-0544 | CB844978 | 0 | 73 | 27 | 0 | 1.97 | 1.51 |
| MRA-0314 | CB848782 | 0 | 61 | 39 | 0 | 1.65 | 2.18 |
| MRA-0545 | CB848901 | 0 | 66 | 34 | 0 | 1.78 | 1.90 |
| MRA-0572 | CB848921 | 0 | 68 | 32 | 0 | 1.84 | 1.78 |
| MRA-0841 | CB849156 | 0 | 68 | 32 | 0 | 1.84 | 1.78 |
| MRA-0096 | CB850344 | 0 | 69 | 31 | 0 | 1.86 | 1.73 |
| MRA-0121 | CB850363 | 0 | 70 | 30 | 0 | 1.89 | 1.67 |
| MRA-0322 | CB848790 | 0 | 70 | 30 | 0 | 1.89 | 1.67 |
| MRA-0912 | CB849219 | 0 | 70 | 30 | 0 | 1.89 | 1.67 |
| MRA-1033 | CB849321 | 0 | 70 | 30 | 0 | 1.89 | 1.67 |
| MRA-1085 | CB849361 | 0 | 70 | 30 | 0 | 1.89 | 1.67 |
| MRA-0157 | CB850381 | 0 | 71 | 29 | 0 | 1.92 | 1.62 |
| MRA-1648 | CB849895 | 0 | 0 | 1 | 0 | 0 | 0.06 |
| M15E-2659 | CB842026 | 6 | 0 | 0 | 0.14 | 0 | 0 |
| M15E-2778 | CB842141 | 6 | 1 | 1 | 0.14 | 0.03 | 0.06 |
| M2PN-1529 | CB845822 | 0 | 5 | 0 | 0 | 0.14 | 0 |
| M2PN-2316 | CB846570 | 0 | 5 | 0 | 0 | 0.14 | 0 |
| M2PN-2533 | CB846779 | 0 | 5 | 0 | 0 | 0.14 | 0 |
| M2PN-3030 | CB847254 | 0 | 5 | 0 | 0 | 0.14 | 0 |
| MRA-1021 | CB849309 | 0 | 0 | 17 | 0 | 0 | 0.95 |
| MRA-1029 | CB849317 | 0 | 0 | 16 | 0 | 0 | 0.89 |
| MRA-1028 | CB849316 | 0 | 0 | 7 | 0 | 0 | 0.39 |
| MRA-1408 | CB849664 | 0 | 0 | 5 | 0 | 0 | 0.28 |

Number of occurrences of each EST in the three libraries are indicated. Percentage of every EST in each library was estimated as the number of occurrences divided by the total number of high-quality ESTs generated from that library.

on a few selected ESTs. Clone MRA-1648 has two homologous sequences in the mouse dbEST and both of these ESTs were identified in mouse retina. We therefore hypothesized that MRA-1648 is a potential retina-enriched EST. qRT-PCR analysis shows that MRA-1648 is highly enriched in retina, with a four-fold higher expression in retina than brain and lung, and over eight-fold higher expression than other tissues (Figure 5a). Similar qRT-PCR analysis confirmed *in silico* temporal expression profiles of several ESTs obtained from E15.5 eye, PN2 eye and adult retinas (Figure 5b).

## Discussion

EST analysis provides a powerful and rapid means of reconstructing the transcriptome of specific tissues and cell types and for identification of differentially expressed genes. In this study, 11,057 clones were isolated and sequenced, yielding 9,638 high-quality ESTs. The accumulation of low-quality sequences in the first 24 plates of clones sequenced from the M15E library indicated initial technical issues in sequencing, rather than significant differences between this and the other two libraries. We have characterized the transcriptional

**Table 6**

**Retina/eye-enriched genes, excluding crystallin and phototransduction genes**

| Clone ID | Gene description | GenBank |
|---|---|---|
| M15E-4178 | Adult male testis cDNA, RIKEN clone:4930517J16 | AK015815 |
| M15E-2565 | Cat eye syndrome chromosome region, candidate 6 homolog (human) (*Cecr6*) | NM_033567 |
| M15E-2026 | Ciliary neurotrophic factor receptor (*Cntfr*) | NM_016673 |
| M15E-0149 | Major intrinsic protein of eye lens fiber (*Mip*) | NM_008600 |
| M15E-1025 | Silver (*Si*) | NM_021882 |
| M15E-5154 | Sine oculis-related homeobox 6 homolog (*Drosophila*) (*Six6*) | NM_011384 |
| M15E-2513 | SNRPN upstream reading frame (*Snurf*) | NM_033174 |
| M15E-3173 | Testis protein TEX13 (*Tex13*) | AF285576 |
| M15E-1676 | Troponin C, cardiac/slow skeletal (*Tncc*) | NM_009393 |
| M2PN-2684 | Adult male testis cDNA, RIKEN clone:4930579P15 | AK016334 |
| M2PN-0941 | Adult retina cDNA, RIKEN clone:A930007D18 | AK020824 |
| M2PN-4223 | Angiopoietin 4 (*Agpt4*) | NM_009641 |
| M2PN-2236 | Beaded filament structural protein in lens-CP94 (*Bfsp1*) | NM_009751 |
| M2PN-3855 | Calpain 12 (*Capn12*) | NM_021894 |
| M2PN-3442 | Forkhead box containing protein N4 (*Foxn4*) | AF323488 |
| M2PN-1767 | Lymphocyte antigen 96 (*Ly96*) | NM_016923 |
| M2PN-0532 | Myogenic factor 6 (*Myf6*) | NM_008657 |
| M2PN-1826 | Zinc finger protein 97 (*Zfp97*) | NM_011765 |
| M2PN-2535 | Similar to zinc finger protein 97, clone MGC:18740 IMAGE:3986622 | BC011426 |
| MRA-0178 | RIKEN adult male spinal cord cDNA clone A330074E19 | BB192844 |
| MRA-0213 | RIKEN 7 days neonate cerebellum cDNA clone A730095K19 | BB260718 |
| MRA-0392 | RIKEN adult retina cDNA clone A930041F21 | BB282979 |
| MRA-0624 | RIKEN 16 days embryo head cDNA clone C130040H24 | BB368057 |
| MRA-0365 | RIKEN 12 days embryo eyeball cDNA clone D230035C09 | BB470735 |
| MRA-0334 | RIKEN adult retina cDNA clone A930002B01 | BB591375 |
| MRA-2067 | 13 Days embryo head cDNA, RIKEN clone:3110045I18 | AK014181 |
| MRA-0913 | 6-Phosphofructo-2-kinase/fructose-2,6-biphosphatase 2 (*Pfkfb2*) | NM_008825 |
| MRA-1771 | Adult male hippocampus cDNA, RIKEN clone:2900046P06 | AK013659 |
| MRA-1770 | Adult male testis cDNA, RIKEN clone:1700110E17 | AK018950 |
| MRA-1154 | Adult male testis cDNA, RIKEN clone:4930438A20 | AK015332 |
| MRA-1937 | Adult retina cDNA, RIKEN clone:A930004D18 | AK020807 |
| MRA-1280 | Adult retina cDNA, RIKEN clone:A930029D14:guanine nucleotide binding protein (G protein), gamma 1 subunit | AK020903 |
| MRA-0210 | cGMP-gated cation channel protein | M84742 |
| MRA-0623 | TWEAK mRNA, partial cds | AF030100 |
| MRA-1629 | *Mus musculus*, clone MGC:28867 IMAGE:4512579 | BC017153 |
| MRA-1185 | *Mus musculus*, clone MGC:38855 IMAGE:5361063 | BC023051 |

profiles of mouse E15.5 eye, PN2 eye and adult retina, by further annotation and analyses of 4,144, 3,701 and 1,793 ESTs, respectively. About half of the identified high-quality ESTs represented known genes. As ESTs have previously been generated from several adult retina libraries, the higher number of known genes in our MRA library was not unexpected. Until the recent large-scale transcription analyses of E14.5 retina [28], ESTs have not been generated from embryonic eyes. Not surprisingly, 29% of ESTs from the M15E library corresponded to novel transcripts, suggesting that EST generation remains a useful approach for gene discovery. Despite the large number of random cDNA clones analyzed, we did not observe much redundancy. Only 0.4% of the ESTs were detected six or more times, not considering mitochondrial and ribosomal genes (7-10%). Approximately 51% of M15E, 58% of M2PN and 67% of MRA ESTs are identified only once

**Table 7**

**Retina/eye-enriched unknown ESTs**

| Clone ID | Accession number | Number of eye EST matches | Total EST match | % Eye EST |
|---|---|---|---|---|
| M15E-0732 | CB840441 | 4 | 4 | 100 |
| M15E-3645 | CB842968 | 162 | 239 | 68 |
| M15E-4248 | CB843444 | 136 | 210 | 65 |
| M15E-4840 | CB843901 | 6 | 7 | 86 |
| M2PN-0590 | CB845024 | 56 | 66 | 85 |
| M2PN-1110 | CB845520 | 6 | 6 | 100 |
| M2PN-1211 | CB845617 | 10 | 13 | 77 |
| M2PN-2132 | CB846398 | 116 | 214 | 54 |
| M2PN-4105 | CB848283 | 197 | 230 | 86 |
| MRA-0135 | CB850369 | 8 | 8 | 100 |
| MRA-0274 | CB848744 | 21 | 21 | 100 |
| MRA-0291 | CB848760 | 16 | 22 | 73 |
| MRA-0304 | CB848772 | 5 | 7 | 71 |
| MRA-0376 | CB848841 | 7 | 11 | 64 |
| MRA-0671 | CB849009 | 7 | 10 | 70 |
| MRA-0729 | CB849066 | 15 | 22 | 68 |
| MRA-1155 | CB849425 | 13 | 21 | 62 |
| MRA-1259 | CB849524 | 4 | 4 | 100 |
| MRA-1408 | CB849664 | 7 | 10 | 70 |
| MRA-1648 | CB849895 | 2 | 2 | 100 |
| MRA-1656 | CB849903 | 4 | 4 | 100 |
| MRA-1706 | CB849949 | 14 | 18 | 78 |
| MRA-1722 | CB849965 | 4 | 4 | 100 |
| MRA-1792 | CB850035 | 9 | 12 | 75 |
| MRA-1858 | CB850097 | 6 | 6 | 100 |

Homology ESTs of each clone were obtained from mouse dbEST and designated as 'eye EST match' if originally isolated from ocular tissue library. % Eye EST was calculated by dividing the number of homology ESTs from ocular tissue by the total number of homologs.

(singletons). This suggests that additional sequencing or *in silico* retinal EST mining may be desired for a more comprehensive transcriptional profiling, especially of the adult retina. It should be noted that the true number of unique clusters will almost certainly be lower since in the absence of full-length sequences it is not always obvious whether two ESTs represent non-overlapping segments of the same gene or two different genes. This issue is minimized by obtaining an average 500 bp or longer sequence from the 5' end, given that the coding regions of most transcripts are less than 1,000 bp in length [1].

Functional categorization of ESTs with known-gene matches underlines general differences in expression profiles of different tissues at distinct developmental stages. The data presented here suggest global changes in gene expression during eye development. For instance, the developing eye exhibited more transcripts involved in cell structure, gene regulation and protein expression, whereas the adult retina had more transcripts representing phototransduction and metabolism. Although our data are limited by the fact that the M15E and M2PN libraries were constructed from whole eye while MRA was from retina only, the overall findings are in good agreement with the expected roles of these tissues at corresponding stages. Similar to results from E14.5 retinal cDNA sets [28], the most abundant transcripts in developing eyes are those of translation factors, house-keeping genes and cell structure/cytoskeletal proteins. Both the M15E and M2PN libraries have higher numbers of clones encoding elongation factor Tu, ubiquitin B, *Gapd* and several structural proteins. Excluding crystallins, hemoglobin alpha adult chain 1 transcript is the most abundant gene in the M15E library. This may reflect active neovascularization of embryonic eyes and dynamic transcriptional activity in the nuclei of immature red blood cells. Levels of hemoglobin transcripts drop dramatically to only six copies in the M2PN library and zero in MRA, perhaps reflecting the gradual maturation of red blood cells and the vascular system. Adult retina plays an important role in vision, as revealed by high expression of phototransduction genes. Although deeper sequencing of these libraries may change the frequencies of individual genes, the overall patterns of gene expression should remain invariant.

In addition to the comparison of global patterns of gene expression, ESTs were utilized to construct *in silico* temporal transcriptional profiles of individual genes in E15.5, PN2 eyes and adult retinas. Relative expression of a single gene in each library was estimated from the number of corresponding ESTs representing this gene normalized by the total number of high-quality ESTs. This approach assumes random sequencing and that the level of activity of a given gene may be inferred from the number of corresponding ESTs obtained. It is intrinsically limited in detecting posttranscriptional regulatory effects and is insufficient in identifying underexpressed genes. However, similar issues are shared by other methods for estimating gene expression, including microarray hybridization. Furthermore, high numbers of analyzed ESTs would undoubtedly increase the sensitivity of this approach. Cross-comparison between our three cDNA libraries identified a number of genes showing a restricted temporal expression. For instance, phototransduction genes are highly and restrictively expressed in the adult retina, whereas crystallin transcript levels are greatly elevated in the E15.5 and PN2 eyes. In addition to providing a list of differentially regulated genes, this approach has enabled us to infer the potential function of unknown ESTs based on their temporal expression patterns. ESTs showing striking differences in the level of expression in the three libraries may play significant roles in eye development or function of mature retina. *In silico* temporal expression profiles should also provide valuable insights into the cellular function of individual genes and may become a useful guide for gene discovery.

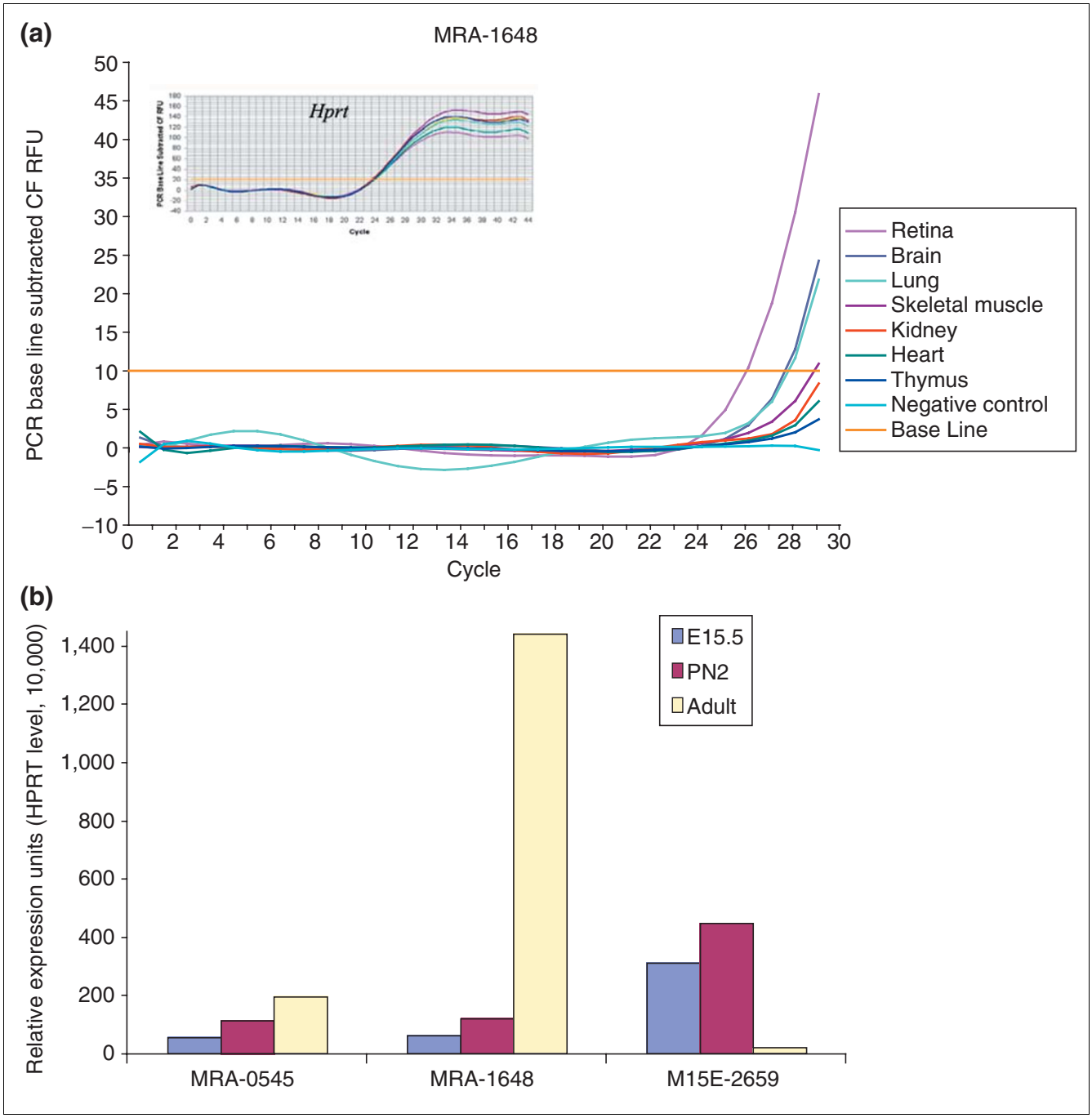



**Figure 5**
qRT-PCR validation of *in silico* expression profiles. **(a)** Clone MRA-1648 showed a four-fold higher expression in retina than in brain or lung. Its expression in other tissues is at least eight-fold less than in the retina. One negative control curve is shown to demonstrate that samples are free of genomic contamination. The inset figure shows the qRT-PCR curves of *Hprt*, all of which cross the baseline at 24 ± 0.5 cycles, indicating equal input of cDNA in different tissue samples. **(b)** The relative expression units of three clones, MRA-0545, MRA-1648 and M15E-2659, in E15.5 eyes (blue), PN2 eyes (purple) and adult retina (yellow) were calculated by their cycle differences from *Hprt* during qRT-PCR experiments. *Hprt* was arbitrarily assigned a value of 10,000. The high transcript level of MRA-0545 and MRA-1648 and low level of M15E-2659 in adult retina confirmed the *in silico* profiles shown in Table 5.

It is believed that a tissue selectively expresses a specific set of genes based on its functional needs. Such preferentially expressed genes are generally of importance and may have disease-causing effects. Mutations in almost all genes that are preferentially expressed in the retina or during eye development have been associated with eye or retinal diseases. Identification of eye or retina-enriched genes is therefore a promising approach to discovering potential disease genes.

This is strengthened by the fact that over half of the retina-enriched genes from the MRA library are known phototransduction genes. Similarly, functional annotation of a number of crystallin genes revealed a restricted expression in ocular tissues including eye or retina. Cystallins have been long known to be major protein components of lens. Recent studies have advocated their role as molecular chaperones [37,38] and similar function is predicted in the retina (R.F. and A.S., unpublished observations). We have validated by qRT-PCR the retina-enriched expression of an EST (MRA-1648) that has been identified only from retina libraries. Our list of known genes or unknown ESTs occurring preferentially in the eye libraries may provide valuable information for gene discovery.

We have identified the human orthologs of 47% of known ESTs identified from our EST set. The chromosomal distribution of ocular genes differs significantly from the observed distribution of 30,000 human genes reported previously [35]. A higher density of ocular genes was found in chromosomes 17 and X, confirming earlier studies that indicated an over-representation of retinal disease loci on these chromosomes [8,18]. Chromosomes 4 and 10 were estimated to have lower ocular gene density. A similar trend was demonstrated in a previous mapping of 3,152 genes from adult human retina [18]. It is noteworthy that our mapping information is highly dependent on the accuracy of UniGene data, and increasing the number of mapped genes will definitely add to the power of this test. We have also observed 277 unique genes localized within the chromosomal intervals of mapped genetic loci for ocular diseases. These genes therefore may be considered as valid candidates for mutation screening.

The identification of ESTs will greatly facilitate in the investigation of the complete transcriptome of specific tissues. ESTs annotated in this study will be a valuable complement to currently archived sequences from ocular tissues and should facilitate eye transcriptome analyses. A comprehensive collection of retina/eye ESTs is an important genomic resource for the positional cloning of disease genes, for large-scale expression studies and for other functional genomic studies. Mouse cDNA microarrays containing over 6,000 retina/eye ESTs have been generated in our laboratory for the transcriptional analysis of ocular tissues during development and from knockout and transgenic mice or mutants [30,39]. The high-level annotation of eye and retinal ESTs, presented here, will greatly facilitate *in silico* expression profiling and experimental approaches utilizing slide microarrays of eye-expressed genes.

## Materials and methods

### Library construction and cDNA sequencing

cDNA libraries were constructed from E15.5 eyes, PN2.5 eyes and adult retina, as previously described [30]. Plasmid clones were randomly selected, and colonies were inoculated into individual wells of 96-well plates containing 175 μL LB media, covered with Breathe-Easy tape (ISC Bioexpress, Kaysville, UT, USA), and incubated at 37°C for 18-22 h. Frozen glycerol stocks were prepared by adding 77 μL of 50% glycerol to each well, and the plate was stored at -80°C. Double-stranded cDNAs were obtained for sequencing either by miniprep (CONCERT Rapid Plasmid Miniprep System, Invitrogen, Carlsbad, CA, USA) or PCR amplification directly from frozen glycerol stocks, as described [30]. DNA sequencing from the 5' end of the cDNA insert was carried out using T7 primer with a high-throughput automated sequencer (Applied Biosystems Inc, Foster City, CA, USA) using standard protocols.

### EST analysis and gene annotation

Raw sequences were first subjected to RepeatMasker [40] and the repeat-masked sequence was used to query NCBI nr database with the BLAST algorithm (National Center for Biotechnology Information, Bethesda, MD, USA) [41]. Sequences matching to nr database entries with an E-value of e-100 or less were classified as a positive BLAST result and RefSeq entries were preferentially selected where available since these have the greatest amount of annotation linked to them. Special consideration was given to BLAST results where our query matched a target DNA/genomic sequence over successive regions with E-values between e-50 and e-99 as this could represent an mRNA matching to different exons of the same gene. Further functional annotation of BLAST positive genes was performed using Perl/Bioperl scripts. In brief, accession numbers were utilized to query the Entrez nucleotide database [32] and UniGene database [33] for information including gene name, gene symbol, UniGene Cluster ID, LocusID, chromosome location and cDNA sources. LocusID was then utilized to query the LocusLink database [34] for gene ontology information, and to search the human UniGene database [33] for human homolog maps.

DNA sequences with no significant matches to NCBI nr database were further BLAST-searched against the mouse subset of dbEST [32]. Sequences matching to dbEST entries with E-value of e-60 or less were considered as positive matches. The cDNA source tissues of all dbEST entries matching the sequence were obtained from the Entrez nucleotide database using Perl scripts.

To assess the redundancy of our clone sets, ESTs with known gene matches were grouped based on identical accession numbers. Unknown and novel ESTs were clustered using default parameters by the NCBI BLASTCLUST, a BLAST score-based single-linkage clustering script [42]. Each cDNA sequence was also BLAST-searched against the collection of all sequences from the M15E, M2PN and MRA libraries using Standalone BLAST [42]. Homology sequences with E-value of e-60 or less were considered to be overlapping sequences belonging to one gene/EST. The quality of each sequence was examined according to objective criteria, such as sequence length and percentage of Ns and As in the sequence. Integrity

of each sequence was eventually determined based on these parameters, with necessary manual analysis. Basically, sequences shorter than 200 bp, with over 30% of Ns or over 50% of As were considered as low quality. Manual analysis was applied to sequences longer than 500 bp but with high-percentage of Ns or As.

### Quantitative real-time PCR (qRT-PCR)

Total RNA was isolated using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) and treated with RQ1 RNase-Free DNase (Promega, Madison, WI, USA) to remove genomic DNA contamination. First-strand cDNA (+*RT* sample) was synthesized by reverse transcription of 2.5 μg of the total RNA using oligo-d(T) primers. Negative control sample (-*RT*) was obtained by incubating 2.5 μg of the total RNA from the same pool without reverse transcriptase. PCR primers were designed using the PRIMER3 software [43]. qRT-PCR was performed in an iCycler IQ real-time PCR Detection system (Bio-Rad, Hercules, CA, USA), and the thermal cycling condition was 3 minutes at 95°C, followed by 45 cycles of 95°C for 30 seconds, 57°C for 30 seconds and 72°C for 30 seconds. SYBR Green (Molecular Probes, Eugene, OR, USA) was added into each reaction for the detection of fluorescence during amplification. PCR reactions from both +*RT* and -*RT* samples were performed in triplicate, and control reactions of *Hprt* were performed on each template to normalize the amount of cDNA present in each reaction. All reaction products were verified with melt curve analysis and agarose gel electrophoresis.

## Additional data file

A complete list of the candidate ocular disease genes identified in the three mouse retina/eye cDNA libraries can be found in (Additional data file 1) with the online version of this article.

## Acknowledgements

We thank Alan J Mears, Mohammad I Othman, Shigeo Yoshida, Sepideh Zareparsi and Yong Zeng for constructive discussions and Sharyn Ferrara for her administrative assistance. This research was supported by grants from the National Institutes of Health (EY11115, including administrative supplements, EY07961, EY10321, EY08123, and core EY07003), Elmer and Sylvia Sramek Foundation, The Foundation Fighting Blindness, Macula Vision Research Foundation, and Research to Prevent Blindness (RPB). A.S. is a recipient of the Lew R Wasserman Merit Award and W.B. of the Senior Investigator Award, both from RPB.

## References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.*: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
3. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al.*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
4. Swaroop A, Zack DJ: **Transcriptome analysis of the retina.** *Genome Biol* 2002, **3:**reviews1022.1-1022.4.
5. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270:**484-487.
6. Cheng G, Porter JD: **Transcriptional profile of rat extraocular muscle by serial analysis of gene expression.** *Invest Ophthalmol Vis Sci* 2002, **43:**1048-1058.
7. Jasper H, Benes V, Atzberger A, Sauer S, Ansorge W, Bohmann D: **A genomic switch at the transition from cell proliferation to terminal differentiation in the *Drosophila* eye.** *Dev Cell* 2002, **3:**511-521.
8. Blackshaw S, Fraioli RE, Furukawa T, Cepko CL: **Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes.** *Cell* 2001, **107:**579-589.
9. Sharon D, Blackshaw S, Cepko CL, Dryja TP: **Profile of the genes expressed in the human peripheral retina, macula, and retinal pigment epithelium determined through serial analysis of gene expression (SAGE).** *Proc Natl Acad Sci USA* 2002, **99:**315-320.
10. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, *et al.*: **Complementary DNA sequencing: expressed sequence tags and human genome project.** *Science* 1991, **252:**1651-1656.
11. Strachan T, Abitbol M, Davidson D, Beckmann JS: **A new dimension for the human genome project: towards comprehensive expression maps.** *Nat Genet* 1997, **16:**126-132.
12. Kawamoto S, Yoshii J, Mizuno K, Ito K, Miyamoto Y, Ohnishi T, Matoba R, Hori N, Matsumoto Y, Okumura T, *et al.*: **BodyMap: a collection of 3' ESTs for analysis of human gene expression information.** *Genome Res* 2000, **10:**1817-1827.
13. Konno H, Fukunishi Y, Shibata K, Itoh M, Carninci P, Sugahara Y, Hayashizaki Y: **Computer-based methods for the mouse full-length cDNA encyclopedia: real-time sequence clustering for construction of a nonredundant cDNA library.** *Genome Res* 2001, **11:**281-289.
14. VanBuren V, Piao Y, Dudekula DB, Qian Y, Carter MG, Martin PR, Stagg CA, Bassey UC, Aiba K, Hamatani T, *et al.*: **Assembly, verification, and initial annotation of the NIA mouse 7.4 K cDNA clone set.** *Genome Res* 2002, **12:**1999-2003.
15. Kan Z, States D, Gish W: **Selecting for functional alternative splices in ESTs.** *Genome Res* 2002, **12:**1837-1845.
16. Xu Q, Modrek B, Lee C: **Genome-wide detection of tissue-specific alternative splicing in the human transcriptome.** *Nucleic Acids Res* 2002, **30:**3754-3766.
17. Sohocki MM, Malone KA, Sullivan LS, Daiger SP: **Localization of retina/pineal-expressed sequences: identification of novel candidate genes for inherited retinal disorders.** *Genomics* 1999, **58:**29-33.
18. Bortoluzzi S, d'Alessi F, Danieli GA: **A novel resource for the study of genes expressed in the adult human retina.** *Invest Ophthalmol Vis Sci* 2000, **41:**3305-3308.
19. Katsanis N, Worley KC, Gonzalez G, Ansley SJ, Lupski JR: **A computational/functional genomics approach for the enrichment of the retinal transcriptome and the identification of positional candidate retinopathy genes.** *Proc Natl Acad Sci USA* 2002, **99:**14326-14331.
20. Bernstein SL, Borst DE, Neuder ME, Wong P: **Characterization of a human fovea cDNA library and regional differential gene expression in the human retina.** *Genomics* 1996, **32:**301-308.
21. Buraczynska M, Mears AJ, Zareparsi S, Farjo R, Filippova E, Yuan Y, MacNee SP, Hughes B, Swaroop A: **Gene expression profile of native human retinal pigment epithelium.** *Invest Ophthalmol Vis Sci* 2002, **43:**603-607.
22. Maubaret C, Delettre C, Sola S, Hamel CP: **Identification of preferentially expressed mRNAs in retina and cochlea.** *DNA Cell Biol* 2002, **21:**781-791.
23. Gieser L, Swaroop A: **Expressed sequence tags and chromosomal localization of cDNA clones from a subtracted retinal pigment epithelium library.** *Genomics* 1992, **13:**873-876.
24. Sinha S, Sharma A, Agarwal N, Swaroop A, Yang-Feng TL: **Expression profile and chromosomal location of cDNA clones, identified from an enriched adult retina library.** *Invest Ophthalmol Vis Sci* 2000, **41:**24-28.
25. Bernstein SL, Borst DE, Wong PW: **Isolation of differentially expressed human fovea genes: candidates for macular disease.** *Mol Vis* 1995, **1:**4.
26. Wang Y, Macke JP, Abella BS, Andreasson K, Worley P, Gilbert DJ, Copeland NG, Jenkins NA, Nathans J: **A large family of putative transmembrane receptors homologous to the product of the *Drosophila* tissue polarity gene *frizzled*.** *J Biol Chem* 1996,

**271:**4468-4476.
27. Shimizu-Matsumoto A, Adachi W, Mizuno K, Inazawa J, Nishida K, Kinoshita S, Matsubara K, Okubo K: **An expression profile of genes in human retina and isolation of a complementary DNA for a novel rod photoreceptor protein.** *Invest Ophthalmol Vis Sci* 1997, **38:**2576-2585.
28. Mu X, Zhao S, Pershad R, Hsieh TF, Scarpa A, Wang SW, White RA, Beremand PD, Thomas TL, Gan L, *et al.*: **Gene expression in the developing mouse retina by EST sequencing and microarray analysis.** *Nucleic Acids Res* 2001, **29:**4983-4993.
29. Wistow G: **A project for ocular bioinformatics: NEIBank.** *Mol Vis* 2002, **8:**161-163.
30. Farjo R, Yu J, Othman MI, Yoshida S, Sheth S, Glaser T, Baehr W, Swaroop A: **Mouse eye gene microarrays for investigating ocular development and disease.** *Vision Res* 2002, **42:**463-470.
31. **I-GENE Database** 2002 [http://www.umich.edu/~igene].
32. **NCBI nucleotide sequence databases** 2002 [http:www.ncbi.nlm.nih.gov/blast/html/blastcgi help.html#nucleotide_databases].
33. **UniGene** 2002 [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene].
34. **LocusLink** 2002 [http://www.ncbi.nlm.nih.gov/LocusLink/].
35. Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, Hui L, Matise TC, McKusick KB, Beckmann JS, *et al.*: **A physical map of 30,000 human genes.** *Science* 1998, **282:**744-746.
36. **RetNet™ Retinal Information Network** 1998 [http://www.sph.uth.tmc.edu/Retnet/].
37. Graw J, Loster J: **Developmental genetics in ophthalmology.** *Ophthalmic Genet* 2003, **24:**1-33.
38. Kurita R, Sagara H, Aoki Y, Link BA, Arai K, Watanabe S: **Suppression of lens growth by alphaA-crystallin promoter-driven expression of diphtheria toxin results in disruption of retinal cell organization in zebrafish.** *Dev Biol* 2003, **255:**113-127.
39. Yu J, Othman MI, Farjo R, Zareparsi S, MacNee SP, Yoshida S, Swaroop A: **Evaluation and optimization of procedures for target labeling and hybridization of cDNA microarrays.** *Mol Vis* 2002, **8:**130-137.
40. **RepeatMasker Web Server** 2002 [http://ftp.genome.washington.edu/cgi-bin/RepeatMasker].
41. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
42. **BLAST download** 1990 [ftp://ftp.ncbi.nih.gov/blast/executables/].
43. **Primer 3** 1990 [http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi].