Genomics 99 (2012) 44–51

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Genome-wide transcriptome analysis in murine neural retina using high-throughput RNA sequencing

Ece D. Gamsiz, Qing Ouyang, Michael Schmidt, Shailender Nagpal, Eric M. Morrow *

Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, 70 Ship Street, Box G-E4, Providence, RI 02903, USA

ARTICLE INFO

Article history:
Received 6 June 2011
Accepted 21 September 2011
Available online 1 October 2011

Keywords:
Retina
Transcriptome
RNA-seq
Mouse
Alternative splicing

ABSTRACT

Genome-wide characterization of the retinal transcriptome is central to understanding development, physiology and disorders of the visual system. Massively parallel, short-read sequencing of mRNA libraries was used to generate an extensive map of the transcriptome of the adult, murine neural retina. RNA-seq data strongly corroborates prior transcriptome studies by microarray and SAGE. However, several novel features of the retinal transcriptome were discovered. For example, retinal disease genes were discovered to be among the most highly expressed in the transcriptome. We also demonstrate other interesting features of the retinal transcriptome, for example, that the retina appears to employ a very specific and restricted set of synaptic vesicle genes, and also that there is persistence of expression of a majority of "neurodevelopmental" genes into adulthood. Retina transcriptome studies utilizing novel sequencing methods have been highly informative and these data may also serve as a resource for the community of researchers.

1. Introduction

The neural retina is an excellent model for studies in neurogenetic disease [1–4]. There are over 200 mapped loci for genetic eye diseases for which simple Mendelian mutations have been found in approximately 170 genes (http://www.sph.uth.tmc.edu/retnet/sum-dis.htm). Genetic mutations lead to visual impairment or blindness in most, and such patients may be characterized functionally (by visual acuity), electrophysiologically (by electroretinogram), and also morphologically through fundoscopic exam. Further still, the cell types of the neural retina are exquisitely well characterized morphologically and biochemically [5]. The approachable layered structure of the neural retina provides critical advantages for experimental studies of genetic eye disease. Finally, in a majority of situations mouse models of human neurogenetic diseases have been highly informative [6]. Thereby, the neural retina specifically offers an important opportunity to examine some of the fundamental features of neurogenetic disease.

Novel, massively-parallel, high-throughput sequencing technologies provide an opportunity for genome-wide observations of the transcriptional architecture of retina genes in a fashion which has not be been previously attainable. The transcriptome is the complete set of transcripts in a cell at a specific stage or under given physiological condition [7]. Genome-wide characterization of the retinal transcriptome is central to understanding cell development, physiology and disease [2,8,9]. High-throughput mRNA sequencing (RNA-Seq) allows simultaneous transcript discovery and abundance estimation [10]. Until recently, microarray platforms (hybridization-based technologies) as well as Serial Analysis of Gene Expression (SAGE) have been used, each with distinct strengths and weaknesses [8,9,11–13]. RNA-Seq utilizing next generation sequencing technology (NGS) has overcome many of the challenges of the previous technologies [14,15]. NGS data have a wide dynamic range of transcript expression for quantification and identification of low abundance transcripts. Furthermore, RNA-Seq discovers all transcripts present in the library within the constraints of the depth of coverage. With microarray technologies, transcripts will only be detected based on prior knowledge as required for probe placement [16]. Also, the size of transcripts may be accurately measured by RNA-seq, as opposed to array hybridization which does not provide information on transcript size.

In this study, we have applied high-throughput RNA sequencing to all mRNAs from the murine neural retina. These data may serve as a resource for the community interested in gene expression in the neural retina. In addition, our analyses have revealed fundamental aspects of the transcriptional architecture of disease genes. We clearly demonstrate that genes that are associated with retinal disease are among the most highly expressed and largest in the transcriptome. We also present other features of interest regarding the transcriptome including that the retina appears to utilize a distinct subset of synaptic vesicle genes, and also, that there is a persistence of expression of many "neurodevelopmental" genes into adulthood. The data

* Corresponding author at: Laboratory for Molecular Medicine, Box G-E4, Brown University, 70 Ship Street, Providence, Rhode Island, 02903 USA. Fax: +1 401 863 9653.

E-mail address: eric_morrow@brown.edu (E.M. Morrow).

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.09.003

presented support the promise of RNA sequencing on NGS platforms for in depth study of gene expression.

2. Results

2.1. Genome-wide transcript profiling in mouse neural retina by RNA-seq

To gain a genome-wide view of gene expression in the neural retina, high-throughput RNA sequencing (RNA-seq) was performed on adult neural retina dissected free from other ocular tissue at postnatal day 21. Two distinct cDNA libraries were prepared to assure biologic replication. Each library was derived from poly(A) + transcripts pooled within one litter of mouse pups from the CD1 line acquired from Charles River Laboratories. Each library was constructed from pooling 8 neural retinae dissected free of other ocular tissue. Libraries were sequenced on the Illumina GAIIx instrument by paired-end chemistry with results as described in Table 1. A total of 49,305,140 reads were obtained for library EMA and 50,818,317 for library EMB. Coverage plots of the five most abundant genes in addition to the transcription factor *Crx* are shown in Figs. 1A–F. Quality scores for each lane of sequence were plotted separately and all sequence was determined to be of high quality with scores above 32 (Supplementary Fig. 1). A transcriptome analysis path was adopted as shown in Fig. 2. Sequenced reads were aligned to the *Mus musculus* (July 2007 NCBI37/mm9) using Bowtie [17]. Splice junctions were then identified using TopHat [18] and transcript abundances were calculated using Cufflinks [10] (Fig. 2). Gene annotations were derived from the Aceview database (Mouse Jun07) [19]. For EMA, in total 45.2 million reads were aligned to the genome with 40 million mapping to unique locations, representing >40× sequence coverage of the 60 Mb *M. musculus* transcriptome. Data for EMB were similar (Table 1). Aceview represents a well established model of gene annotation for mouse and is among the most comprehensive set of gene models available. Other annotations include RefSeq, Vega, Ensembl, Gene Trap, Geneid, MGC and Genescan are available from the UCSC browser (www.genome.ucsc.edu). For purposes of completeness, we compared an analysis across these two databases. Concordance for expression levels when reads were aligned to Aceview and UCSC gene models was high and plotted with R-squared values of over 0.97 for EMA (Supplementary Fig. 2a). Comparison of expression levels obtained for EMB when mapped to the alternative reference libraries showed a similar concordance of findings with R-squared values of 0.94 (data not shown).

Comparison of results of the two libraries, i.e. biologic replicates, EMA and EMB, also showed a high degree of concordance. As described, each library was sequenced to a similar depth of coverage (Table 1). Utilizing the analysis path as shown (Fig. 2), 29,580 transcripts were identified in EMA and 31,685 transcripts were identified in EMB. 70% of all transcripts identified were discovered in both EMA and EMB, and expression levels showed high concordance (R-squared of 0.86) across trials (Supplementary Fig. 2b). Transcripts which we discovered in only one of EMA or EMB were dropped from general analyses described below. In total, 15,251 known genes and 20,558 transcripts were coordinately identified using RNA-Seq in our two libraries from mouse retina. Two or more isoforms were found in 3655 genes. The complete list of genes with alternative splicing, transcript sizes and abundances is included in Supplementary Table 1.

Table 1
RNA sequencing in murine neural retina.

Library	Read length (bp)	Reads produced (million)	Sequenced bases (Mb)	Aligned reads (millions)	Reads aligned uniquely (millions)	Unique reads overlapping exons (%)
EMA	60	49.3	2958.3	45.2	40.0	83.1
EMB	60	50.8	3049.1	45.2	41.7	80.7

2.2. Transcripts associated with disease are among the most abundant and largest in the transcriptome

Gene abundances were quantified in *F*ragments *P*er *K*ilobase of exon model per *M*illion mapped fragments (FPKM) and exhibited a range between 0.125 and 2481.5 (Table 2) [10]. Based on transcripts discovered in both replicates the mean and median expression levels were 10.07 FPKM and 4.11 FPKM respectively. The twenty most abundantly expressed genes are listed in Table 3. Genes related with eye diseases were taken from Retinal Information Network (RetNet) (http://www.sph.uth.tmc.edu/retnet/sum-dis.htm) which list 159 genes in total associated with retinal disease as of 12/5/2010. Across both biologic replicates, transcripts for 114 genes associated with retinal disease were identified (127 in EMA and 132 in EMB), and many of these genes were among the most abundantly expressed. For example, the three genes with highest expression in both replicates were three retinal disease genes: rhodopsin (Rho) (2481.5 +/− 232.01 FPKM) is associated with dominant retinitis pigmentosa, dominant congenital stationary and night blindness; guanine nucleotide-binding protein G(t) subunit alpha-1 (Gnat1) (also known as the transducin alpha-1 chain) (2092.48 +/− 281.38 FPKM) is associated with dominant congenital stationary and night blindness; and retinal S-antigen (Sag) (1046.75 +/− 393.9 FPKM) is associated with recessive Oguchi disease and recessive retinitis pigmentosa (Table 4). Further still, while disease genes represent less than 1% of the genes in the transcriptome (114 total disease genes of 15251 genes), 30% (6 of 20) of the 20 most highly expressed genes were associated with disease (Figs. 3A, B and Table 3). Statistical analysis indicates clearly that the expression level of disease genes is significantly higher than those of non-disease genes (Chi-Square 72.07, DF = 1, $p<0.0001$). The gene expression scores demonstrated a profound skewness calculated as 40.79 using the univariate procedure; however, the statistically significance was strongly maintained when alternative, non-parametric test procedures were utilized which do not assume a normalized data distribution such as the Wilcoxon Rank Sums/Kruskal–Wallis (Chi-Square 72.07, DF = 1, $p<0.0001$).

Of the 159 disease genes identified in the RetNet database (http://www.sph.uth.tmc.edu/retnet/sum-dis.htm), 126 were identified concordantly in our RNA-seq experiments (Supplementary Table 2). Six genes were discovered only in one of the two replicates. The 33 genes which were not identified are listed in Supplementary Table 3. Of these 33 genes, 18 genes which were not identified in neural retina had cogent rationale for the absence. For example, ten of these genes are known to be expressed exclusively in retinal pigment epithelium, lens, vitreous, vascular elements or exclusively during development (i.e. are not expressed in adult neural retina). (Recall that our tissue sample was neural retina dissected free of other ocular tissue.) Several (an additional three) clearly function at a systemic level with extra-ocular gene expression, for example, complement factor H (Cfh) or *Abcc6* which are secreted by non-neuronal cells. Finally, five of the known disease genes which were not identified were strict mitochondrial expressed genes.

We sought to identify differences between the most abundantly expressed disease genes and those which were not abundantly expressed. We performed pathway analysis on the most abundant and least abundant twenty genes using Gene Ontology Enrichment Analysis Software Toolkit (GOEAST)[20]. There were several similarities between the most highly expressed disease genes as compared to the least abundantly expressed disease genes (Supplementary Tables 4 and 5 as well as Supplementary Figs. 3 and 4 for most abundant and least abundant twenty disease genes, respectively). Among the common features were biological processes which included sensory perception, development of retinal structures and retinol metabolic processes. Also in common were cellular processes which involved photoreceptor outer segment, inner segment and connecting cilium. Differences included in the top group phototransduction ($p<2.46e-10$) which involved four of the top twenty genes, nucleotide metabolism such as

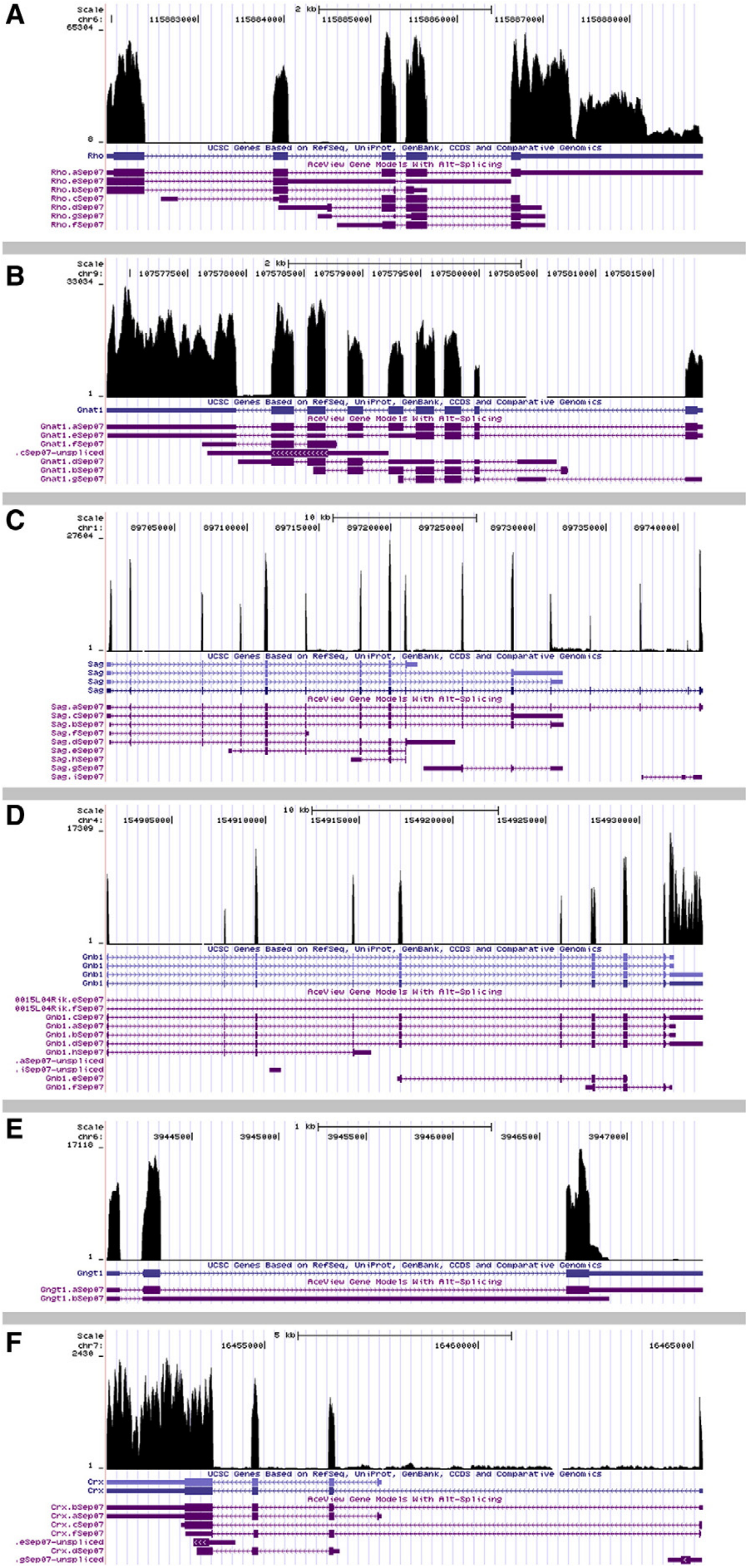

Fig. 1. Coverage plots of most abundant genes (A) *Rho*, (B) *Gnat1*, (C) *Sag*, (D) *Gnb1*, (E) *Gngt1* and (F) *Crx*.

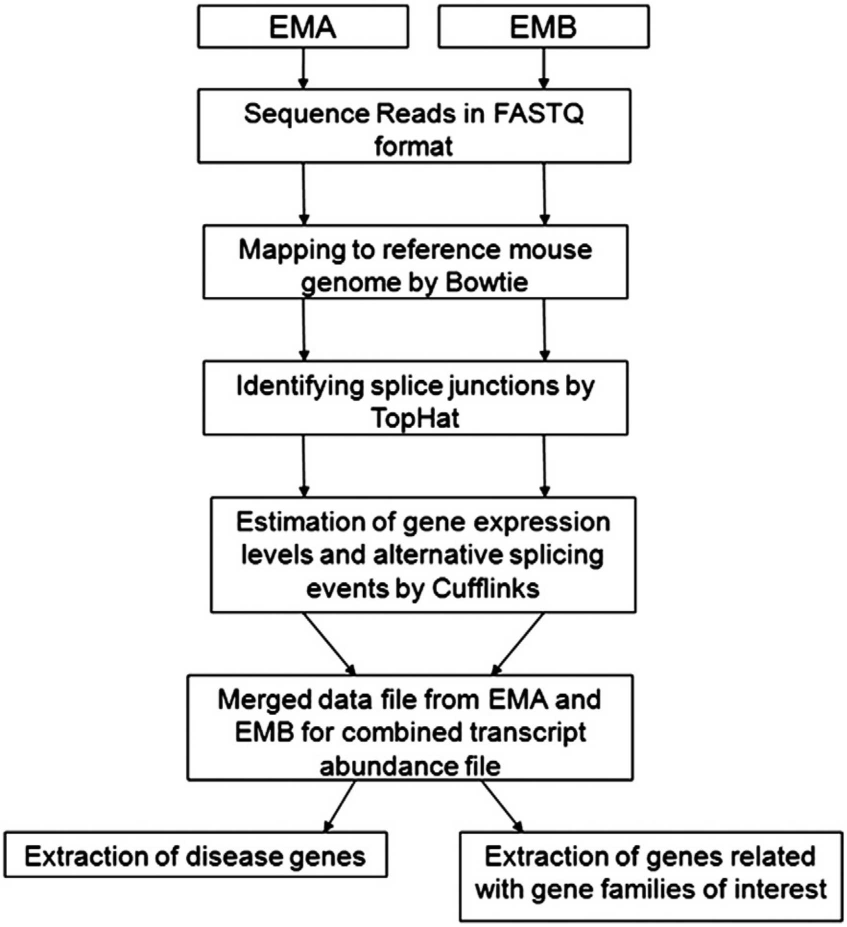
EMA
EMB
Sequence Reads in FASTQ format
Mapping to reference mouse genome by Bowtie
Identifying splice junctions by TopHat
Estimation of gene expression levels and alternative splicing events by Cufflinks
Merged data file from EMA and EMB for combined transcript abundance file
Extraction of disease genes
Extraction of genes related with gene families of interest

Fig. 2. The data analysis steps in the computational pipeline to analyze RNA-Seq data from mouse retina to obtain expressed known transcripts with expression levels and number of isoforms as well as to extract disease genes.

cyclic-nucleotide phosphodiesterase activity (p<8.57e−8) which included three genes, and rod cell development (p<1.8e−4) although this last category only involved a single gene of the twenty. Pathways which emerged for the disease genes that were least abundantly expressed which were not shared with the top genes, included molecular mediator of immune response (p<2.29e−4) involving two genes, extracellular matrix genes (p<7.7e−4) involving three genes, and cone cell differentiation (p<2.79e−4) although this only involved a single gene.

On average, transcripts which have been implicated in disease were also larger than non-disease transcripts and had more alternative splicing events. For example, transcripts associated with disease were found to be 4333.4 kb on average as compared to 3323.1 kb for non-disease transcripts (Chi-Square 13.3638, DF = 1, p = 3e−4). Further still, we noticed that transcripts associated with disease showed a small but statistically significant increase in alternative transcripts (number of transcripts were 1.55 and 1.35 for disease and non-disease genes, respectively) (Chi-Square 7.4753, DF = 1, p = 6.3e−3). To address the possibility that the increased transcript size and increased transcript number were correlated with transcript expression level, we conducted correlation studies of these factors. Across the full dataset, we did not identify a correlation between transcript abundance and size (Pearson's coefficient = 0.0007), nor transcript size and alternative transcripts (Pearson's coefficient = −0.093), nor abundance and alternative transcripts (Pearson's coefficient = −0.036). Thereby, the increase in abundance, size and transcript number of genes associated with disease appears to be a novel and fundamental feature of disease genes in the retina transcriptome.

Table 2
Summary analysis of transcript expression and splicing.

	Range	Mean	Median
Gene expression (FPKM)	0.125–2481.5	10.07	4.11
Number of transcripts/gene	1–8	1.35	1

Table 3
Most highly expressed known genes from Aceview database.

Gene	Number of isoforms	Average expression (FPKM)
Rho	1	2481.5 ± 164.1
Gnat1	2	2092.5 ± 199
Sag	2	1046.7 ± 278.5
Gnb1	2	625.5 ± 158.5
Gngt1	1	560.7 ± 73.5
Hcn1	1	546.8 ± 165.5
Loc218963	1	482.8 ± 36.1
Rasana	1	409.9 ± 100.2
Jarid2	4	386.3 ± 19 .0
Yukowara	1	380.1 ± 97 .0
Rbp3	1	375.5 ± 13.8
Pde6g	1	370.4 ± 237.8
Prph2	1	367.9 ± 70.8
Pdc	1	345.1 ± 78.9
Cpe	1	340 ± 6.7
Seyjaw	1	334.8 ± 132.8
Hexb	1	321.1 ± 309.9
Snap25	3	320.7 ± 27.5
Surera	1	296.0 ± 91.6
Syp	1	276.0 ± 25.1

Table 4
Most highly expressed known genes related with eye diseases.

Gene	Isoforms	Average Expression (FPKM)	Disease
Rho	Rho.a	2481.5 ± 164.1	Retinitis pigmentosa, congenital stationary night blindness
Gnat1	Gnat1.g	2092.5 ± 199.0	Congenital stationary night blindness
	Gnat1.a	136.9 ± 136.3	
Sag	Sag.h	1046.7 ± 278.5	Oguchi disease, retinitis pigmentosa
	Sag.i	214.8 ± 36.3	
Rbp3	Rbp3.a	375.5 ± 13.8	Retinitis pigmentosa
Pde6g	Pde6g.b	370.4 ± 237.8	Retinitis pigmentosa
Prph2	Prph2.a	367.9 ± 70.8	Retinitis pigmentosa, macular dystrophy, adult vitelliform macular dystrophy
Pde6a	Pde6a.a	210.8 ± 0.8	Retinitis pigmentosa
	Pde6a.c	106.2 ± 27.0	
Pde6b	Pde6b.h	169.0 ± 143.1	Retinitis pigmentosa, congenital Stationary night blindness
	Pde6b.a	58.1 ± 2.1	
	Pde6b.b	4.8 ± 0.0	
Rdh12	Rdh12.b	156.1 ± 3.5	Leber congenital amaurosis with severe childhood retinal dystrophy
Prom1	Prom1.o	143.4 ± 13.0	Macular dystrophy, bull's-eye, retinitis Pigmentosa with macular degeneration, Stargardt-like macular dystrophy, Macular dystrophy, bull's-eye, cone rod dystrophy
	Prom1.k	12.2 ± 3.6	
	Prom1.m	0.7 ± 0.2	
	Prom1.l	0.4 ± 0.2	

2.3. Cell-type specific gene expression by comparison to transcriptome studies from previous methods

With regard to cell-type specific gene expression, we compared our dataset to previously published transcriptome studies. For example, 43 genes had previously been identified by SAGE to be rod-enriched [8]. We identified 42 of these previously identified genes in our dataset. The average transcript level for these genes is 227.0 FPKM with the range from 1.3 to 2481.5 FPKM (Supplementary Table 6). With regard to bipolar cells, we identified 71/73 previously identified genes [21]. The average transcript level for these bipolar genes is 21.5 FPKM with the range from 0.3 to 125.2 FPKM. For amacrine cells, 86/88 previously identified genes were found to be expressed in our dataset [21]. The average amacrine cell transcript level for these genes was 33.3 FPKM, with a range from 0.3 to 432.0 FPKM. For ganglion cells, we identified expression in 47/49 of the ganglion cell enriched genes previously reported [22], and the range of gene expression was from 0.2 to 126.9 FPKM with an average expression level of 24.4 FPKM. For Muller cells, gene expression was noted in 69/70 genes noted to have highest Muller expression [23]. Average gene expression in Mullers was 39.0 FPKM and ranged from 0.5 to 226.5 FPKM. Failure to find the very small minority of genes described is attributed to nomenclature discrepancies, i.e. gene names that were not identified in AceView and could not be otherwise identified. In the case of three genes only (*Hmgcs1* and *Hub1* (Zfp282) in ganglion cells and *Prss2* in Mullers), genes noted to be expressed in previous studies were found not to be expressed in our mouse dataset. For the ganglion cell genes, this might reflect species differences. Of interest, some genes which may be predicted to have very low gene expression were discovered in our dataset. For example, *Opsin4* (Opn4, melanopsin) is only expressed in a small fraction (1%) of all ganglion cells (which represent a small fraction of retinal cells) is detected in both of our trials at levels of 1.3 and 1.4 FPKM. Finally, to verify that there was not an overabundance of contamination from retinal pigment epithelium, seven signature RPE genes were found to have no expression or very little expression (0.4 FPKM in only one trial). Our data are therefore consistent with very little contamination from RPE.

2.4. Alternative splicing observed in 24% of genes

We assessed the extent of alternative splicing and identified 3655 genes with more than two identical isoforms expressed in both replicates. The number of alternative transcripts per gene was observed to

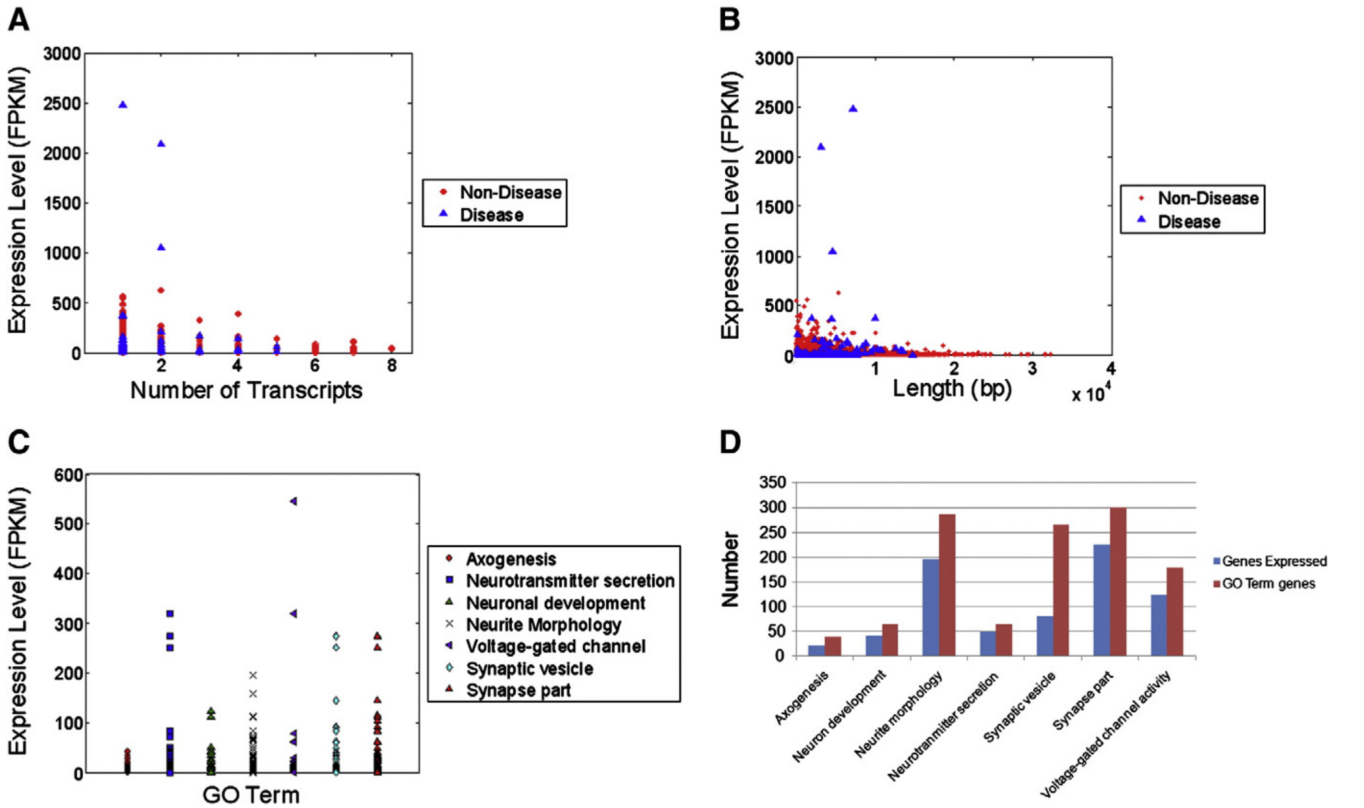

Fig. 3. Comparison of disease and non-disease genes expressed in EMA and EMB. (A) Expression levels in FPKM versus transcript number, (B) expression levels in FPKM versus transcript length (bp), (C) Analysis of gene expression level (FPKM) based on gene category, (D) Histogram of number of genes expressed in given GO Term category.

be between 1 and 8. Using conservative measures (the requirement that the isoform be identified in both replicates), we identified alternative splicing in 24% of genes. The average abundance of the most highly expressed transcript for genes with multiple isoforms was 12 FPKM. As indicated, we did not discover a correlation of number of isoforms with transcript abundance or size. Genes with the greatest number of transcripts are shown in Table 5. For example, *Abca8a* has 8 isoforms in both replicates and the expression of isoforms ranges between 18.4 and 41.9 FPKM.

In order to compare our alternative transcript results with the prior published results, we curated all data from a Pubmed search on alternative splicing of retina genes as described in the Materials and methods section. This search on January 14, 2011 returned 144 references. The studies of alternative splicing in the literature were performed in a variety of tissues (i.e. including for example retinal pigmented epithelium), species and developmental stages. In situations wherein the tissue, species, or developmental stages were similar, the number of transcripts discovered in our study generally agreed or exceeded published reports (Supplementary Table 7). For example, *Crx*, *Six3* and *Nrxn2* were found in our study to have the same number of transcripts as has been determined previously in the literature. For other genes with similar experimental conditions, our study identified more isoforms (concordantly in both biologic replicates), for example, the RNA-seq experiments identified four transcripts for *Rpgrip1* while a prior report identified only two transcripts [24]. Similarly, for *Cabp1*, RNA-seq identified five transcripts while two were previously documented in the literature [25]. The complete list of transcripts discovered (including those discovered in both EMA and EMB) is found in Supplementary Tables 8, 9 and 10. We validated a test set of approximately thirty transcripts by PCR using exons specific primers. In a majority of transcripts (greater than 85%), the alternative transcripts were identified by PCR (data not shown). For example, in the case of gene *Rpgrip1* and *Cabp1*, wherein transcript numbers exceeded previously reported, all transcripts tested were validated (three of three for *Rpgrip1* and four of four tested for *Cabp1*) which confirms these new predictions of transcripts in excess of prior reports. Similarly, for gene *Atp1a1*, which was one of the most alternatively spliced genes, seven of seven transcripts tested were validated by PCR testing.

Table 5
Genes with greatest number of alternative transcripts.

Gene	Number of isoforms	Isoforms	Average expression level (FPKM)
Abca8a	8	*Abca8a.a*	42.0 ± 0.5
		Abca8a.f	37.3 ± 0.9
		Abca8a.e	35.9 ± 3.9
		Abca8a.b	30.2 ± 1.9
		Abca8a.g	28.8 ± 0.1
		Abca8a.d	22.4 ± 0.6
		Abca8a.h-unspliced	19.1 ± 0.7
		Abca8a.c	18.4 ± 3.4
Med15	7	*Med15.o*	104.9 ± 11.8
		Med15.d	21.5 ± 0.4
		Med15.a	11.0 ± 10.1
		Med15.n	7.8 ± 0.0
		Med15.s	3.4 ± 0.1
		Med15.u	2.5 ± 0.1
		Med15.i	0.8 ± 0.0
Atp1a1	7	*Atp1a1.a*	49.9 ± 1.2
		Atp1a1.i-unspliced	42.4 ± 1.3
		Atp1a1.b	39.2 ± 0.9
		Atp1a1.f	37.4 ± 11.9
		Atp1a1.j-unspliced	34.4 ± 0.3
		Atp1a1.h-unspliced	25.8 ± 4.1
		Atp1a1.e-unspliced	14.5 ± 4.6
Wdr1	7	*Wdr1.g*	35.3 ± 0.5
		Wdr1.a	33.9 ± 2.5
		Wdr1.h	24.4 ± 0.8
		Wdr1.k	17.3 ± 1.1
		Wdr1.i	15.8 ± 3.4
		Wdr1.j	15.0 ± 2.5
		Wdr1.l	9.4 ± 3.7
Anapc1	7	*Anapc1.a*	22.3 ± 0.3
		Anapc1.f	19.0 ± 1.6
		Anapc1.j	14.7 ± 0.0
		Anapc1.k	12.2 ± 4.5
		Anapc1.l	6.8 ± 4.6
		Anapc1.b	5.7 ± 2.1
		Anapc1.e	2.6 ± 0.5

2.5. Retina uses a distinct subset of synaptic vesicle genes

Genome-wide analysis of transcription using RNA-seq also revealed several novel features of the retinal transcriptome. We studied genes in several gene ontology (GO) groups such as axonogenesis, neuron development, neurite morphology, neurotransmitter secretion, synaptic vesicles, synapse part and voltage-gated channel activity. The representation of transcripts within these GO categories is shown in Figs. 3C and D. Interestingly, despite the derivation of the transcriptome from adult tissue 42 out of 68 neuron development genes (greater than 60%) were expressed in both samples. Expression level of those genes ranges between 0.5 and 123 FPKM (Supplementary Table 11). Average expression level of neuron development genes is 18 FPKM. Also 195 genes of 285 neurite morphology genes (greater than 65%) are expressed adult in both samples and the average expression level is 14.9 FPKM (Supplementary Table 12). Due to the fundamental role of voltage-gated ion channels in neuronal physiology, genes related with this family are expected to be highly expressed in retina (Supplementary Table 13). A surprisingly large percentage of these genes, 124 of 179 genes (nearly 70%) defined to be in this family, were expressed in both of replicates and at high levels. The most abundant gene in this family and among all gene families chosen for our study is *Hcn1*, Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 1 gene and the average expression level for voltage-gated ion channels was 16.7 FPKM. Synaptic vesicles have critical roles in neurotransmission in a neuron. It was unanticipated that only 81 out of 265 genes (approximately 30%) related with synaptic vesicles were discovered. In spite of the low number of genes expressed in this family, the average of expression level is quite high (28.4 FPKM) (Supplementary Table 14). This result supports a very interesting model: the retina utilizes a relatively specific and restricted subset of synaptic vesicle genes.

3. Discussion

We have generated a comprehensive map of retina transcriptome in mouse using high throughput, paired-end RNA-seq. We are making these data publically available and anticipate this could be a highly accessed resource for researchers interested in gene expression in the neural retina. There have been several transcriptome studies in retina, all were done by using previous technologies such as microarray and serial analysis of gene expression (SAGE) [8, 9]. Dorrell et al. reported 2635 known genes in a microarray study on developing mouse retina [26]. In this study, postnatal mouse retina development was studied including the postnatal day 21 which is the same sampling day as our study. Using SAGE, Blackshaw et al. reported gene expression during retinal development, but focused on genes expressed by mammalian rods [8]. A microarray study done by Livesey et al. focused on transcriptional network controlled by *Crx* using cDNA microarray [27]. Zhang et al. also used cDNA microarray to study the mouse retinal development from embryonic day 12.5 to postnatal day 21 [28]. In the present study, we used RNA-seq to accurately identify transcript abundances (with a broad dynamic range), size and alternative splicing information in mouse retina at postnatal day 21. Our study has revealed several novel features of the retina transcriptome, including in particular the architecture of known retinal disease genes. We identified 15,251 known genes expressed concordantly with precise alternative splicing information and 3655 of those were found to have more than 2 isoforms.

The neural retina has a large number of known disease genes which act in a simple Mendelian fashion (greater than 150 genes) which lead to visual impairment and blindness (http://www.sph.uth.tmc.edu/retnet/sum-dis.htm). Few other tissues have such a well characterized and highly specific signature of disease genes. These attributes of the neural retina make this tissue an excellent model to study the transcriptional architecture of neurogenetic disease by genome-wide RNA-seq. Here we have discovered a number of features of the transcriptional architecture of disease genes including disease genes which are among the most abundantly expressed, and are generally larger with more alternative transcripts than non-disease genes. It is possible that the high level of expression of disease genes may be a feature of the retina, but we propose that these attributes may alternatively be fundamental properties of disease genes. One possible interpretation with regard to abundance may be that disease genes are commonly expressed in rod photoreceptors and these cells are among the most abundant cell types in the murine neural retina, representing greater than 70% of cells (Lavail 1980; Morrow 1998). While this observation may explain part of the high abundance of disease genes, the increase in alternative splicing and the larger size of disease-related genes seem likely independent characteristics of the disease gene transcriptional architecture. More work will be necessary in other neural tissues to examine this result in greater detail.

Our data provide interesting findings on alternative splicing in mouse retina. Alternative splicing occurs in 3655 genes (24%) of the genes identified in both libraries with a large number of genes demonstrating over two transcripts. Our findings are highly consistent with prior estimates of alternatively splicing in human retina where alternative splicing has been estimated to be in 26% of retina genes in a study based on expressed sequence tags (ESTs) [29]. Alternative splicing events in our study showed high concordance in terms of gene abundances ($R^2 = 0.85$, data not shown) yet we anticipate that our results may represent an underestimate of alternative splicing given our strict criteria for identifying transcripts in our biologic replicates. However, we have provided the full analysis of the conservative list of transcripts (discovered in all replicates) and also all transcripts from individual replicates in Supplementary data.

Finally, genome-wide profiling of retinal transcriptome has revealed other interesting features of retina biology. For example, we noted that of all the synaptic vesicle genes in the genome, only 30% of such genes are expressed in the retina; that is, the retina uses a very specific signature of synaptic vesicle genes. Further work in different neural tissues will likely elucidate distinct signatures which may suggest specific physiologic properties of neurons in different neuronal circuits. In addition, the observation that neuronal development genes show persistent expression in adult tissues is also compelling. This argues that a majority of such genes may have various functions including in neuronal development and potentially in the maintenance of neuronal health.

In summary, this study presents the most comprehensive view of the transcriptome of the murine neural retina to date using novel, massively-parallel sequencing technologies. Our analysis uses the state-of-the-art tools for calculations of transcript abundance, size and alternative splicing. These data are provided as a resource for the community of researchers studying gene expression in retina.

4. Materials and methods

4.1. Sample collection, library preparation and sequencing

The neural retina was dissected free of any ocular tissue from 21 day old CD1 mice which were purchased from Charles River Laboratories. Poly (A) + transcripts were isolated by following the manufacturer's instructions (*mir*Vana™ miRNA Isolation Kit, Ambion®) from dissected retina. EMA and EMB libraries were constructed from duplicate samples based on established protocols by Illumina (mRNA Sequencing Sample Preparation Guide Cat # RS-930-1001). Isolated cDNA libraries were sequenced by paired-end chemistry via Illumina Genome Analyzer IIx. On average 50 million of 60 bp reads were obtained from each library.

4.2. Data analysis

Sequencing reads in FASTQ format were mapped to build 37.1 of the mouse genome as well as splice junctions were identified using Bowtie version 0.12.5.0 [17] and TopHat version 1.1.0 [18]. Average library size and read length were 280 bp and 60 bp paired-end reads, respectively. Transcript abundances were estimated by using Cufflinks in Fragments per kilobase of transcript per million fragments sequenced (FPKM) [10]. All transcripts identified by Cufflinks version 0.9.1 were matched to the gene annotations taken from Aceview database [19] by Cuffcompare which is a part of Cufflinks. Parameters used to run TopHat and Cufflinks are shown in Supplementary Table 18. The program pipeline used for the data analysis is shown in Fig. 2. Output of this pipeline contains files with gene annotations, transcripts, abundances and loci of all transcripts. By using custom-made scripts, these files were further analyzed to extract eye disease genes as well as gene families of interest.

FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) was used to check the quality of raw sequence data. Each lane (EMA s_5, EMA s_6, EMB s_7 and EMB s_8) was analyzed with FastQC and results were shown in Supplementary Figs. 1A–D. According to FastQC quality scores, raw sequencing reads of our samples are high quality.

Coverage plots were prepared by using Bedtools program (http://code.google.com/p/bedtools/). Firstly, the "accepted_hits.bam" file which was generated by TopHat was converted into a BEDgraph file by Bedtools program. BEDgraph file was converted into BigWig file which was used for visualization on UCSC browser (http://genome.ucsc.edu/cgi-bin/hgGateway).

4.3. Curation of alternative splicing genes

Pubmed search was performed to search for genes with alternative splicing isoforms between 1/14/2010 and 1/17/2010. Keywords such as “alternative splicing of retina genes, alternative splicing isoforms in retina” were entered to Pubmed and the abstracts of first 100 results were read. If the corresponding gene in the abstract was expressed in our samples and the abstract indicated experimental findings on alternative splicing isoforms, other sections of the article were examined. Alternative transcripts with solid experimental support were included in Supplementary Table 7.

4.4. Statistics

Correlation between transcript abundance, size and alternative transcripts was calculated by Pearson's correlation on Excel. For the comparison of disease and non-disease genes in terms of expression level, number of alternative transcripts and transcript length, NPAR1WAY procedure in SAS was used.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2011.09.003.

Acknowledgments

EMM holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund. EDG has been supported by a postdoctoral fellowship award from the Center for Vision Research at Brown University. The authors would like to thank Robert Stout PhD for advice with statistical analysis, Lingsheng Dong MD for assistance with bioinformatics, and members of the Morrow lab for helpful discussion. We are grateful to David Berson PhD and Anita Zimmerman PhD, for providing feedback on the manuscript. Aspects of this work were presented in a poster at the Annual Association for Research in Vision and Ophthalmology (ARVO) 2011 meeting on May 3, 2011 in Fort Lauderdale. Additional sources of funding include NIMH K23 MH080954-05 (EMM), NIH S10 RR027634-01, NIH P30 RR031153-01A1 and R NIH P20 RR018728-10.

References

[1] C.L. Cepko, Genomics approaches to photoreceptor development and disease, Harvey Lect 97 (2001) 85–110.

[2] A. Swaroop, D.J. Zack, Transcriptome analysis of the retina, Genome Biol. 3 (2002) REVIEWS1022.

[3] C.L. Montana, J.C. Corbo, Inherited diseases of photoreceptors and prospects for gene therapy, Pharmacogenomics 9 (2008) 335–347.

[4] A.N. Bramall, A.F. Wright, S.G. Jacobson, R.R. McInnes, The genomic, biochemical, and cellular responses of the retina in inherited photoreceptor degenerations and prospects for the treatment of these disorders, Annu. Rev. Neurosci. 33 (2010) 441–472.

[5] R.H. Masland, Neuronal cell types, Curr. Biol. 14 (2004) R497–R500.

[6] E.L. Fletcher, A.I. Jobling, K.A. Vessey, C. Luu, R.H. Guymer, P.N. Baird, Animal models of retinal disease, Prog. Mol. Biol. Transl. Sci. 100 (2011) 211–286.

[7] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63.

[8] S. Blackshaw, R.E. Fraioli, T. Furukawa, C.L. Cepko, Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes, Cell 107 (2001) 579–589.

[9] A.S. Hackam, J. Qian, D.M. Liu, T. Gunatilaka, R.H. Farkas, I. Chowers, M. Kageyama, G. Parmigiani, D.J. Zack, Comparative gene expression analysis of murine retina and brain, Mol. Vis. 10 (2004) 637–649.

[10] C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, L. Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nat. Biotechnol. 28 (2010) 511–515.

[11] D. Sharon, S. Blackshaw, C.L. Cepko, T.P. Dryja, 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 (2002) 315–320.

[12] I. Chowers, T.L. Gunatilaka, R.H. Farkas, J. Qian, A.S. Hackam, E. Duh, M. Kageyama, C.W. Wang, A. Vora, P.A. Campochiaro, D.J. Zack, Identification of novel genes preferentially expressed in the retina using a custom human retina cDNA microarray, Investig. Ophthalmol. Vis. Sci. 44 (2003) 3732–3741.

[13] D.A. Lamba, T.A. Reh, Microarray characterization of human embryonic stem cell-derived retinal cultures, Investig. Ophthalmol. Vis. Sci. 52 (2011) 4897–4906.

[14] V. Costa, C. Angelini, I. De Feis, A. Ciccodicola, Uncovering the complexity of transcriptomes with RNA-Seq, J. Biomed. Biotechnol. 2010 (2010) 1–19.

[15] A. Mortazavi, B.A. Williams, K. Mccue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq, Nat. Methods 5 (2008) 621–628.

[16] F. Ozsolak, P.M. Milos, RNA sequencing: advances, challenges and opportunities, Nat. Rev. Genet. 12 (2011) 87–98.

[17] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (2009).

[18] C. Trapnell, L. Pachter, S.L. Salzberg, TopHat: discovering splice junctions with RNA-Seq, Bioinformatics 25 (2009) 1105–1111.

[19] D. Thierry-Mieg, J. Thierry-Mieg, AceView: a comprehensive cDNA-supported gene and transcripts annotation, Genome Biol. 7 (Suppl 1) (2006) 1–14 S12.

[20] Q. Zheng, X.J. Wang, GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis, Nucleic Acids Res. 36 (2008) W358–W363.

[21] S. Blackshaw, S. Harpavat, J. Trimarchi, L. Cai, H.Y. Huang, W.P. Kuo, G. Weber, K. Lee, R.E. Fraioli, S.H. Cho, R. Yung, E. Asch, L. Ohno-Machado, W.H. Wong, C.L. Cepko, Genomic analysis of mouse retinal development, Plos Biol. 2 (2004) 1411–1431.

[22] D. Ivanov, G. Dvoriantchikova, L. Nathanson, S.J. McKinnon, V.I. Shestopalov, Microarray analysis of gene expression in adult retinal ganglion cells, FEBS Lett. 580 (2006) 331–335.

[23] K. Roesch, A.P. Jadhav, J.M. Trimarchi, M.B. Stadler, B. Roska, B.B. Sun, C.L. Cepko, The transcriptome of retinal miller glial cells, J. Comp. Neurol. 509 (2008) 225–238.

[24] X. Lu, P.A. Ferreira, Identification of novel murine- and human-specific RPGRIP1 splice variants with distinct expression profiles and subcellular localization, Investig. Ophthalmol. Vis. Sci. 46 (2005) 1882–1890.

[25] F. Haeseleer, I. Sokal, C.L. Verlinde, H. Erdjument-Bromage, P. Tempst, A.N. Pronin, J.L. Benovic, R.N. Fariss, K. Palczewski, Five members of a novel Ca(2+)-binding protein (CABP) subfamily with similarity to calmodulin, J. Biol. Chem. 275 (2000) 1247–1260.

[26] M.I. Dorrell, E. Aguilar, C. Weber, M. Friedlander, Global gene expression analysis of the developing postnatal mouse retina, Investig. Ophthalmol. Vis. Sci. 45 (2004) 1009–1019.

[27] F.J. Livesey, T. Furukawa, M.A. Steffen, G.M. Church, C.L. Cepko, Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene Crx, Curr. Biol. 10 (2000) 301–310.

[28] S.S. Zhang, X. Xu, M.G. Liu, H. Zhao, M.B. Soares, C.J. Barnstable, X.Y. Fu, A biphasic pattern of gene expression during mouse retina development, BMC Dev. Biol. 6 (2006) 48.

[29] G. Yeo, D. Holste, G. Kreiman, C.B. Burge, Variation in alternative splicing across human tissues, Genome Biol. 5 (2004).