BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 18 2010, pages 2289–2297
doi:10.1093/bioinformatics/btq408

Systems biology Advance Access publication July 8, 2010

Computational analysis of tissue-specific gene networks: application to murine retinal functional studies

Jianfei Hu[1], Jun Wan[1], Laszlo Hackler Jr.[1], Donald J. Zack[1,2,3,4] and Jiang Qian[1,*]

[1]Wilmer Institute, [2]Department of Molecular Biology and Genetics, [3]Department of Neuroscience and [4]McKusick-Nathans Institute of Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA

Associate Editor: Trey Ideker

ABSTRACT

Motivation: The vertebrate retina is a complex neuronal tissue, and its development, normal functioning and response to injury and disease is subject to a variety of genetic factors. To understand better the regulatory and functional relationships between the genes expressed within the retina, we constructed an interactive gene network of the mouse retina by applying a Bayesian statistics approach to information derived from a variety of gene expression, protein–protein interaction and gene ontology annotation databases.

Results: The network contains 673 retina-related genes. Most of them are obtained through manual literature-based curation, while the others are the genes preferentially expressed in the retina. These retina-related genes are linked by 3403 potential functional associations in the network. The prediction on the gene functional association using the Bayesian approach outperforms predictions using only one source of information. The network includes five major gene clusters, each enriched in different biological activities. There are several applications to this network. First, we identified ∼50 hub genes that are predicted to play particularly important roles in the function of the retina. Some of them are not yet well studied. Second, we can predict novel gene functions using ‘guilt by association’ method. Third, we also predicted novel retinal disease-associated genes based on the network analysis.

Availability: To provide easy access to the retinal network, we constructed an interactive web tool, named MoReNet, which is available at http://bioinfo.wilmer.jhu.edu/morenet/

Contact: jiang.qian@jhmi.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 14, 2010; revised on June 10, 2010; accepted on July 06, 2010

1 INTRODUCTION

In the post-genome era, a major challenge is to identify the functional elements in the genomes (Birney *et al*., 2007). Both experimental and computational approaches have been successful in identification of many functional elements (Bertone *et al*., 2004; Blanchette *et al*., 2006; Cawley *et al*., 2004; Harrison *et al*., 2005; Kim *et al*., 2005). A next step is the determination of the properties of these elements, especially for the genes that encode proteins. The properties of a gene include the biological function of its gene product, the biological pathways in which it acts, and any disease associations. One traditional method to predict the function of a novel protein is based on homology, as proteins with similar primary sequences or 3D structures tend to share biological function (Bateman *et al*., 2004; Whisstock and Lesk, 2003). To understand the biological pathways in which a protein is involved or to find evidence of a disease association generally requires detailed and time-consuming laboratory work. Traditional methods to study a gene’s properties often focus on the protein of interest, and ignore its cellular context. An emerging technology to predict gene properties takes a different approach, one based on the analysis of gene networks (D’Haeseleer *et al*., 2000).

Gene co-expression networks have been used to predict protein function. DNA microarray studies can generate huge amounts of gene expression data under a variety of different physiological conditions. Based on the assumption that genes with similar gene expression profiles are likely to be functionally related, researchers have been successful in applying correlation-based approaches to predict gene function (Eisen *et al*., 1998; Lee *et al*., 2004a, b; Niehrs and Pollet, 1999; Zhang *et al*., 2004). Similar concepts have also been applied to predict gene function based on protein–protein interaction (PPI) networks (Deng *et al*., 2004; Jansen *et al*., 2003; Letovsky and Kasif, 2003).

A more informative gene network, and one with hopefully greater predictive power, could be constructed by integrating diverse types of biological datasets. For example, probabilistic functional networks have been developed for a number of organisms, and these studies have resulted in the identification of novel interactions (Guan *et al*., 2008; Kim *et al*., 2008; Lee *et al*., 2004a, b; Pena-Castillo *et al*., 2008). Network analysis has also been applied to predict disease-associated genes (Franke *et al*., 2006; Goh *et al*., 2007; Ideker and Sharan, 2008; Lage *et al*., 2007; Oti *et al*., 2006; Pujana *et al*., 2007). However, these studies often ignore the specific behavior of different cell types (or tissues) and as a result yield only a ‘unified network’ that is generic for all tissues. For higher eukaryotes, we believe that important aspects of gene networks are intrinsically tissue-specific. In other words, a ‘unified network’ will likely only reflect the common, basic cellular processes shared by all types of tissues or cell types. In order to better understand the specific biological systems for higher eukaryotic cells, we need to construct tissue-specific gene networks.

In the analysis described here, we used mouse retina as a model system to construct a tissue-specific gene network by integrating different types of available biological data supporting functional relationships of genes. In the network, genes belonging to the

*To whom correspondence should be addressed.

same group are more likely to join in a common pathway, and are thus more likely functionally related with each other. This network can help us decipher the function of unknown genes, learn about new functions for known genes, and also suggest candidate genes potentially responsible for a variety of retinal diseases.

2 MATERIALS AND METHODS

2.1 Microarray data

Nine retina-related microarray datasets, representing retinas under different conditions, were downloaded from the NCBI GEO database (Supplementary Table S1). Also included were a microarray dataset covering 61 mouse tissues (Su *et al*., 2004) so as to be able to describe gene expression profiles across tissues, and transcriptome data from laser capture microdissection (LCM) generated retinal pigmented epithelium (RPE), outer nuclear (photoreceptor), inner nuclear, ganglion cell layer samples derived from adult mouse retina (L. Hackler *et al*., manuscript in preparation; NCBI GEO: GSE19304).

All datasets were log2 transformed. Affymetrix probe identifiers or cDNA identifiers were converted to official gene symbol name. Genes without corresponding official gene symbol names were discarded. When multiple probes correspond to the same gene symbol name, all probes were averaged to obtain its gene expression level. Since none of the microarray datasets used in this study was designed for probing exon expression levels, we cannot differentiate the expression of different transcript variants of the same gene in a systematic way in this study.

We analyzed each set of microarray data separately instead of merging them as one dataset, since a pair of genes might be co-expressed in a specific experiment condition, but not in all, and the signal in one condition might be overwhelmed by noise in other conditions when microarray data are merged (Lee *et al*., 2004a, b). We also analyzed the correlation between the datasets to ensure the independence of each dataset.

2.2 PPI data

PPI data was collected from five databases: DIP (Database of Interacting Proteins, http://dip.doe-mbi.ucla.edu/), MIPS (Mammalian PPI database, http://mips.gsf.de/proj/ppi/), IntAct (ftp://ftp.ebi.ac.uk/pub/databases/intact/current), Biogrid (http://www.thebiogrid.org/) and HPRD (Human Protein Reference Database, http://www.hprd.org). 3328 mouse PPI pairs were obtained. In addition, we also collected 55 048 human PPI pairs. Mouse and human PPI data are highly correlated; if a protein pair interacts in mouse, it is highly likely that it also interacts in human. To reduce redundancy, which would add undue weight in the generation of the network, the overlapped PPI were deleted from the list of human PPI.

2.3 Known biological pathway and Gene Ontology

Known pathway annotations of 6008 mouse genes were downloaded from the KEGG database (ftp://ftp.genome.jp/pub/kegg/genes). Gene Ontology (GO) annotations for 18 184 mouse genes were downloaded from GO.

2.4 Manual curation of retina genes

We performed the literature search for retina-related genes in two steps. First, we ran our script to search Pubmed by ‘retina + gene name’ for all mouse genes. This provided an initial list for retina-related genes. Second, we performed a manual check on these genes to see whether the genes are indeed retina-related with experimental evidence.

2.5 Bayesian integration of data from different sources

According to Bayesian theory,

$$P(AB) = P(A\,|\,B) \times P(B) = P(B\,|\,A) \times P(A). \qquad (1)$$

Thus,

$$P(A|B) = \frac{P(B|A)}{P(B)} \times P(A). \qquad (2)$$

If A represents the functional relationship of a gene pair, and $B = B_1 \ldots B_n$ represent n data supporting functional relationship, such as correlation of GO annotation, gene expression as well as PPIs, then we can get

$$P(true|B_1 \ldots B_n) = \frac{P(B_1 \ldots B_n|true)}{P(B_1 \ldots B_n)} \times P(true). \qquad (3)$$

In the equation above, P(true) is the prior probability that the functional relationship of the gene pair is true, P($B_1 \ldots B_n$) is the joint probability that n supporting data occurs, P($B_1 \ldots B_n$|true) is the joint probability that n supporting data occurs when the functional relationship is true. The later item is difficult to compute since we don’t know all true events. To circumvent this difficulty, we consider another equation simultaneously,

$$P(false|B_1 \ldots B_n) = \frac{P(B_1 \ldots B_n|false)}{P(B_1 \ldots B_n)} \times P(false). \qquad (4)$$

Division of Equation (3) by Equation (4) yields

$$\frac{P(true|B_1 \ldots B_n)}{P(false|B_1 \ldots B_n)} = \frac{P(B_1 \ldots B_n|true)}{P(B_1 \ldots B_n false)} \times \frac{P(true)}{P(false)}. \qquad (5)$$

When n data are independent with each other,

$$\begin{aligned}\frac{P(true|B_1 \ldots B_n)}{P(false|B_1 \ldots B_n)} &= \prod_{i=1}^{n} \frac{P(B_i|true)}{P(B_i|false)} \times \frac{P(true)}{P(false)} \\ &= \prod_{i=1}^{n} L(B_i) \times \frac{P(true)}{P(false)} = L \times \frac{P(true)}{P(false)}.\end{aligned} \qquad (6)$$

Here,

$$L = \prod_{i=1}^{n} L(B_i) = \prod_{i=1}^{n} \frac{P(B_i|pos)}{P(B_i|neg)}. \qquad (7)$$

Equation (7) is easier to use than Equation (3) because L and $L(B_i)$ can be estimated from a small set of known true and false data.

2.6 GO distance score

The average semantic similarity was employed to evaluate the GO distance of two genes (Lord *et al*., 2003). According to the definition, the distance of two GO terms is decided by their most specific common parental term. For example, GO:0004317 and GO:0004730 has six common parental GO terms, the most specific one is GO:0016836, which relates to 38 mouse genes, thus the semantic distance between GO:0004317 and GO:0004730 equals to ln(38) = 3.64. The more similar two GO terms, the smaller is the semantic distance score. The distance of two genes is defined as the average distance score of their related GO terms (Lord *et al*., 2003).

2.7 Statistical significance of cutoff for functional association

As an approach to establish thresholds with statistical significance probability, we reiteratively randomly selected two gene pairs and exchanged their score values. We repeated the procedures as many times as the total number of gene pairs in the data to ensure all gene pairs have been shuffled. In this process, each gene pair is shuffled two times on average. Taking the microarray data containing N genes as an example, we randomly selected two gene pairs to exchange their Pearson Correlation Coefficient (PCC) and repeated the procedure $N \times (N-1)/2$ times. We then computed the combined L-values for each gene pairs to obtain the distribution of L-values [see Equation (7) for definition of L]. The full permutation was repeated 10 times to create the background. From this background distribution of $L-$values we chose a cutoff of $L_{\text{cut}} = 11$, which corresponds to $P = 0.005$. This indicates that, in random condition only 0.5% of gene pairs have L-value > 11.

2.8 Novel function prediction of gene

We predicted the novel function of individual genes by the GO enrichment of its neighbors (genes linked to it) in the network. For a given gene, we took all its neighbors in the network as a gene set, and then computed the enriched GO term for the gene set using the hypergeometric distribution.

$$p(N,M,n,m)=\sum_{i=m}^{n}\frac{C_M^m \times C_{N-M}^{n-m}}{C_N^M}. \tag{8}$$

here, N is the total number of mouse genes with GO annotation and n is the number of neighbor genes of the given gene, M is the number of mouse genes annotated with a given GO term and m is the number of neighbor genes annotated with the GO term.

2.9 Evaluation of disease related gene prediction

The same 'guilt by association' method was applied to predict novel disease-associated genes. To systematically evaluate the power of disease-association genes, we performed a cross-validation analysis. For a group of genes known to be associated to a retinal disease, we randomly selected three-fourth of them as training set and predicted whether other genes in the network (including the remaining known disease genes) are the disease-associated genes. We sorted the predicted genes according to the P-values, and gave each gene a rank index. This procedure was repeated 100 times to get the average ranking for each gene.

2.10 Identification of gene clusters by hierarchical clustering

We first constructed a linkage matrix of the gene pairs. If the L-value of two genes is larger than L_{cut}, the corresponding matrix value is set as L_{cut}, otherwise it is set as 0. We then performed a hierarchical cluster (average linkage) on this matrix using CLUSTER software (Reich *et al*., 2004).

3 RESULTS

3.1 Identification of retina-related genes

We first compiled the parts list for the retinal gene network, the set of retina-related genes. The retina-related genes were obtained from two sources. The first was manually curated retina genes supported by published literature. Our literature search found 512 retina-related genes, each supported by at least one published paper (see Supplementary Table S2 for the genes and the related publications). Second, we also obtained additional retina-related genes based on the gene expression profiles across various tissues. Using the NCBI UniGene dataset, we obtained genes that are preferentially expressed in retina using our previously published approach (enrichment score >8 and P-value <0.05) (Yu *et al*., 2006). By this standard, we identified 256 genes preferentially expressed in retina. These two gene sets were then combined, which resulted in a total of 673 retina-related murine genes.

3.2 Construction of retinal functional network

We attempted to predict functionally related retinal genes by integrating genomic and proteomic information. To train the Bayesian model, we first created positive and negative datasets of functionally related gene pairs based on KEGG pathway information. Functionally related gene pairs were defined as two genes that share the same pathway annotation. By this definition, we obtained 1 755 695 functionally related and 16 289 333 functionally unrelated gene pairs (genes with KEGG annotation not found in the same pathway).

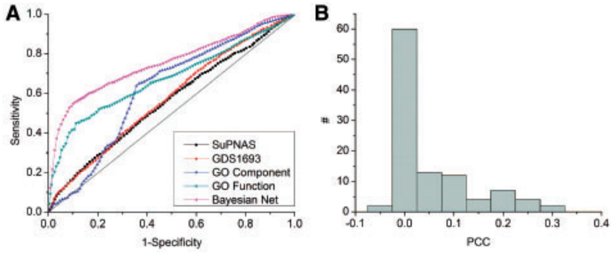

Fig. 1. Performance evaluation and independence analysis of datasets. (**A**) The Bayesian integrated prediction was evaluated against predictions from single dataset by ROC, where positives and negatives are defined as the gene pairs in KEGG pathways that share or do not share co-annotations. The Bayesian network has the best performance. (**B**) PCC of the used datasets. The correlations of the datasets are very low, with the maximum of 0.33, indicating that these datasets are highly independent.

We used three independent data sources to predict functionally related gene pairs: gene expression, GO and PPIs. For microarray data, we computed the PCC of expression pattern for each gene pair. For GO data, we computed the GO similarity score for each gene pair. For PPI data, we divided them into three categories: interaction, ortholog interaction and no interaction. We then calculated the odds ratio (i.e. L) with the observed variable based on positive and negative data sets. For each gene pair, we first computed its L-value in each data set [$L(B_i)$], then obtained its overall L-value by multiplying these L-values [Equation (7)]. The odds ratio represents the likelihood that the gene pairs are functionally related given the observed values.

We selected a conservative L_{cut} of 11, which corresponds to P-value 0.005 based on permutation simulation (see Section 2), to define functionally related gene pairs. Based on an $L_{\text{cut}}=11$ cutoff, 497 genes are linked by 3403 potential functional relationships. This conservative cutoff yielded a high specificity (98.5%) with a sensitivity of 20.2%, suggesting that the network is of high fidelity, although it may not reflect the whole picture of retinal gene network.

3.3 Performance evaluation of prediction and independence analysis of datasets

We evaluated the performance of the Bayesian approach using Receiver Operating Characteristic (ROC) curve. The true positive rate (sensitivity) and false positive rate (1-specificity) were computed for a series of cutoffs. We employed 5-fold cross-validation: we randomly divided both positive and negative data sets into five subsets; one of the subsets is selected as the test set and the other four subsets are together to form a training set. We then evaluated the prediction of functional relationship on the training set using KEGG database as standard. This process was repeated five times. As a comparison, we also show the ROC curves of prediction made from a single type of dataset. The result indicates that the performance of Bayesian method is better than the prediction based on any single source information (Fig. 1A).

To use the naïve Bayesian method, we must ensure that the datasets used are independent. Following a previously reported method, we evaluated the correlation of dataset by PPC of L-values based on different data sets (Lu *et al*., 2005). For each pair of datasets, we first extracted the gene pairs shared by both datasets, and then computed the correlation coefficient of L-values of these gene pairs in the two datasets. Our result shows that 83% of PCCs are <0.15 and the maximum is 0.33, indicating that these datasets are likely to be independent from each other (Fig. 1B).

3.4 Global organization of the network

After we obtained the murine retinal gene network, we first analyzed the global organization of gene network by examining the major gene clusters in the network. Groups of genes that are densely connected to each other in the network may represent functional modules in which the genes are highly related in function and/or cooperate in some biological processes. We performed hierarchical clusters analysis using CLUSTER software (see Section 2), and found five major gene clusters with gene number >10 (five red block in Fig. 2A). The genes in the same cluster are densely connected with each other (Fig. 2B), and GO analysis indicates that these five gene clusters are enriched in certain GO annotation terms (Table 1).

Cluster I is enriched in growth factor genes. Growth factors are molecules that can stimulate the growth and differentiation of cells. There are 24 genes in this cluster, 16 are annotated with growth factor activity (GO:0008083, Bonferroni-corrected $P = 4.21\text{E-}24$). Among them, *Vegfa* has been well known to be able to prevent apoptotic death of retinal endothelial cells and rescues the retinal vasculature (Alon *et al.*, 1995). The remaining eight genes are *Angpt2*, *Cartpt*, *Fasl*, *Shh*, *Slit1*, *Slit2*, *Ttr and Vip*. Since these genes are connected with 16 growth factor genes in the network, we predict that they have growth factor activity or their activities cooperate with growth factors.

Cluster II contains crystallin genes. There are 17 genes in this cluster, 16 of them are annotated with structural constituent of eye lens (GO:0005212, Bonferroni-corrected $P = 9.85\text{E-}45$). Although these genes are previously well known for their role as constituents of the eye lens, a growing body of evidence shows that these genes are also expressed in the retina and they may have vital functions in protecting retinal neurons from damage by environmental and metabolic stress (Wang *et al.*, 2009; Wu *et al.*, 2009; Xi *et al.*, 2003).

Cluster III contains transcription regulating genes. There are 50 genes in this cluster, 49 of them are annotated with transcription regulator activity (GO:0030528, Bonferroni-corrected $P = 1.98\text{E-}56$), including well-known retinal transcription factors *Crx*, *Nr2e3* and *Nrl*. The detailed structure of the module can reveal the cooperative activities among the transcription factors (see later for details).

Cluster IV is enriched in cation transmembrane transporter genes. There are 46 genes in this cluster, 34 of them are annotated with cation transmembrane transporter activity (GO:0008324, Bonferroni-corrected $P = 1.15\text{E-}40$). The proteins encoded by these cation-transport-related genes are important for the transition of visual signal from photon to electric signal in the retina.

One cluster (cluster V) is enriched for visual perception genes. There are 44 genes in this cluster, 33 of them are annotated with both visual perception (GO:0007601, Bonferroni-corrected $P = 3.07\text{E-}65$) and sensory perception of light stimulus (GO:0050953, Bonferroni-corrected $P = 4.78\text{E-}65$). Even among the remaining 11 genes that are not annotated as visual perception genes by the current GO database, some are indeed related to visual perception. *Cplx4* is an essential regulator of transmitter release at retinal ribbon synapses, and its knockout leads to aberrant adjustment of transmitter release at the photoreceptor synapse (Reim *et al.*, 2009). *Gngt1* is a rod photoreceptor transducin protein (Scherer *et al.*, 1996). *Grk1* is involved in signal transduction process and in response to light stimulus (Chen *et al.*, 1999). *Slc24a1* is a potassium-dependent sodium/calcium exchanger in rod photoreceptors (Sharon *et al.*, 2002). *Aipl1*, *Ankrd33* and *Cabp4* are all photoreceptor proteins, and the mutation in *Aipl1* can cause Leber congenital amaurosis, a severe early onset retinopathy that leads to visual impairment in infants (Haeseleer *et al.*, 2004; Liu *et al.*, 2004; Sanuki, *et al.*, 2010; Zeitz *et al.*, 2006).

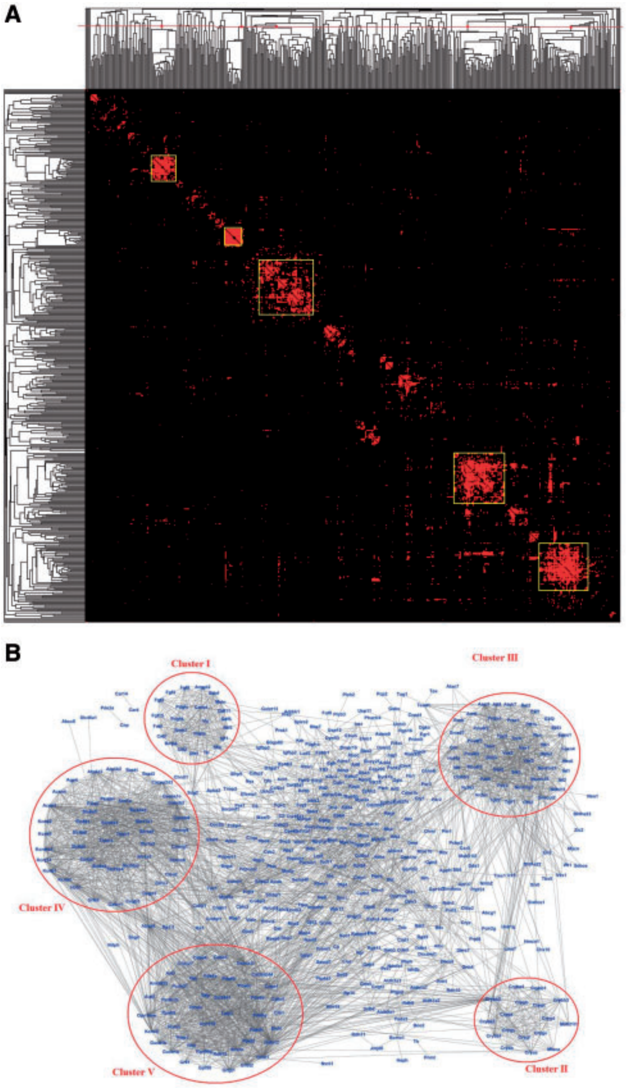

Fig. 2. Hierarchical clustering of retina genes. (**A**) Five gene clusters were identified, each enriched in different functions. (**B**) The functional gene network of the mouse retina ($L_{\text{cut}} = 11$). Four hundred and ninety-seven genes are connected by 3403 edges in this network.

These examples show that the gene network can predict novel biological functions. Furthermore, these five major groups may represent the most important and unique aspects of retinal gene network, which are unlikely to be discovered in a 'unified network'.

Table 1. Five gene clusters in the network

Clusters	Genes	GO enrichment		
		GO number	Term name	*P*
I (24)	*Bdnf*, *Bmp4*, *Bmp7*, *Cntf*, *Fgf15*, *Fgf2*, *Fgf3*, *Fgf8*, *Gdf11*, *Gdf6*, *Igf1*, *Mstn*, *Ntf5*, *Pdgfc*, *Pdgfd*, *Vegfa* *Angpt2*, *Cartpt*, *Fasl*, *Shh*, *Slit1*, *Slit2*, *Ttr*, *Vip*	GO:0008083	Growth factor activity	4.21E-24[a] (16)[b]
II (17)	*Cryaa*, *Cryab*, *Cryba1*, *Cryba2*, *Cryba4*, *Crybb1*, *Crybb2*, *Crybb3*, *Cryga*, *Crygb*, *Crygc*, *Crygd*, *Cryge*, *Crygf*, *Crygs*, *Mlana* *Mab21l1*	GO:0005212	Structural constituent of eye lens	9.85E-45 (16)
III (50)	*Arntl*, *Ascl1*, *Atf4*, *Atoh7*, *Bcl2*, *Crx*, *Ctnnb1*, *E2f1*, *E2f2*, *E2f3*, *Epas1*, *Foxn4*, *Hes6*, *Hipk2*, *Hnf1a*, *Isl1*, *Isl2*, *Neurod1*, *Neurod4*, *Neurog2*, *Notch1*, *Nr2e1*, *Nr2e3*, *Nrl*, *Onecut2*, *Otx2*, *Pax2*, *Pax6*, *Pou4f1*, *Pou4f2*, *Pou4f3*, *Pou6f2*, *Prox1*, *Ptf1a*, *Rax*, *Rb1*, *Rorb*, *Six3*, *Smad7*, *Sox2*, *Tbx2*, *Tbx3*, *Tead1*, *Tgif1*, *Tgif2*, *Thrb*, *Vax1*, *Vax2*, *Vsx2* *Prdm1*	GO:0030528	Transcription regulator activity	1.98E-56 (49)
IV (46)	*Accn1*, *Accn4*, *Atp1a3*, *Cacna1e*, *Cacna2d3*, *Cacnb2*, *Cacnb4*, *Cacng4*, *Cacng7*, *Chrna6*, *Cnga3*, *Cngb3*, *Hcn1*, *Kcna2*, *Kcnb1*, *Kcnj14*, *Kcnma1*, *Kcns1*, *Kcns2*, *P2rx7*, *Pkd2l1*, *Scn1a*, *Slc17a7*, *Slc1a2*, *Slc1a3*, *Slc1a7*, *Slc24a2*, *Slc24a4*, *Slc4a4*, *Slc6a6*, *Slc6a9*, *Slc8a1*, *Slc8a3*, *Trpm3* *Aqp4*, *Atp1b2*, *Best1*, *Best3*, *Clcn2*, *Clcn3*, *Gabrr1*, *Gria4*, *Slc16a8*, *Slc4a7*, *Slco4a1*, *Trpm1*	GO:0008324	Cation transmembrane transporter activity	1.15E-40 (34)
V (44)	*Arr3*, *Cacna1f*, *Cacna2d4*, *Cnga1*, *Crb1*, *Gabrr2*, *Gnat1*, *Gnat2*, *Gpr98*, *Guca1a*, *Guca1b*, *Kcnv2*, *Opn1mw*, *Pcdh15*, *Pcdh21*, *Pdc*, *Pde6a*, *Pde6b*, *Pde6c*, *Pde6d*, *Pde6g*, *Pde6h*, *Ppef2*, *Rcvrn*, *Rdh12*, *Rgs9bp*, *Rho*, *Rlbp1*, *Rom1*, *Rpe65*, *Sag*, *Tulp1*, *Unc119* *Aipl1*, *Ankrd33*,*C79127*,*Cabp4*, *Cabp5*, *Cplx4*, *Drd4*, *Gngt1*, *Grk1*, *Rtbdn*, *Slc24a1*	GO:0007601	Visual perception	3.07E-65 (33)

[a] Bonferroni-corrected *P*-value.
[b] Number of genes in cluster with the GO term annotation.

Our results demonstrate again the importance of constructing tissue specific gene networks.

3.5 Hub genes

In a gene network, the genes with more connections are generally more important for fitness than those with fewer connection, and the proteins encoded by these genes often influence more pathways than the less connected ones (Jeong *et al*., 2001). These highly connected genes are called hub genes. We operationally defined the 10% of genes with the highest degree of connections as hub genes, where the degree of a gene is defined as the number of its linked neighbors. Under this definition, 50 hub genes were found, with degrees ranging from 31 to 75 (Table 2). These hub genes and their neighbors cover 327 genes (65.8%) in the network.

The 50 hub genes, as compared to non-hub genes, do in fact turn out to be more likely to link to more functional modules and participate in more biological processes. Since we did not use the number of GO terms as input in our modeling, as an independent test, we compared the number of GO terms associated with hub and non-hub genes. The average number of molecular function GO terms and biological process GO terms associated to a mouse gene are 8.9 and 21.4, respectively, while the corresponding numbers of GO terms related to a hub gene are 19.7 and 55.1, respectively. One extreme example is *P2rx7*, which has 33 related GO function terms and 321 GO process terms (annotated GO terms and their parental terms). Its role is very diverse, including ion channel activity (GO:0005216), copper and zinc ion binding (GO:0005507, GO:0008270) and protein kinase activity (GO:0043539). Consistent with the central role that this hub gene plays in the network, knock-out of *P2rx7* in mice leads to a wide variety of abnormalities,

Table 2. List of hub genes

Gene	Degree	Gene	Degree	Gene	Degree
Gnat1	75 (956[a])	Rcvrn	48 (380)	Gabrr1	37 (163)
Cacna1f	63 (42)	Crybb2	47 (6)	Slc4a4	36 (6)
Gabrr2	61 (9)	Pdc	47 (17)	Guca1b	36 (52)
Pde6g	56 (15)	Sag	47 (338)	Guca1a	35 (95)
Kcnma1	55 (7)	Gnat2	46 (32)	Ascl1	35 (35)
Accn1	55 (2)	Kcnb1	46 (6)	Sox2	35 (29)
Scn1a	54 (3)	P2rx7	45 (21)	Pde6h	34 (18)
Abca4	54 (164)	Opn1mw	43 (10)	Pde6b	34 (122)
Cacnb2	54 (0)	Crx	42 (176)	Nr2e3	33 (98)
Myo5a	53 (31)	Cabp4	42 (6)	Shh	33 (95)
Cnga1	52 (16)	Unc119	41 (17)	Kcna2	33 (6)
Cacnb4	49 (57)	Hcn1	40 (13)	Atp1a3	33 (0)
Kcnv2	48 (8)	Slc17a7	40 (23)	Myh10	32 (16)
Tulp1	48 (25)	Gngt1	40 (4)	Kcnj14	32 (4)
Pax6	48 (334)	Ctnnb1	39 (91)	Rgs9bp	32 (23)
Ppef2	48 (1)	Cacna2d3	38 (2)	Rho	31 (615)
Ankrd33	48 (0)	Slc24a1	37 (1)		

[a] PubMed hit number by ‘retina and gene name’.

including neural progenitor cell death, reduced bone formation and brain sickness (Delarasse *et al*., 2009; Li *et al*., 2009; Mingam *et al*., 2008).

It has been reported that *Crx*, *Nrl* and *Nr2e3* play a central role in the mammalian photoreceptor transcriptional network (Chen *et al*., 1997, 2005; Hsiau *et al*., 2007; Mears *et al*., 2001; Qian *et al*., 2005). *Crx* is expressed in both rods and cones and activates gene expression in both. *Nrl* and *Nr2e3* are rod-specific and are required for activation of rod genes and repression of cone genes. In our network, *Crx* and *Nr2e3* are both hub genes, with degrees of 42 and 33, respectively. Although *Nrl* is not a hub gene, its degree (25) is also high, belonging to the top 20%. Interestingly, the *L*-values that represent the strength of association between genes are high among these three genes (>330), suggesting that these three factors cooperate tightly in regulating their target genes. This is consistent with experimental data that indicates that they can act together synergistically in activating expression of photoreceptor genes. Furthermore, we found that genes connected to these three genes also enrich in transcription factors. Sixty four genes are predicted to be functionally related with these three genes (Supplementary Fig. S1). Forty of them are annotated with transcription regulator activity (GO:0030528, Bonferroni-corrected $P = 7.04\text{E-}26$) and 33 of them are annotated with transcription factor activity (GO:0003700, Bonferroni-corrected $P = 3.70\text{E-}24$).

One might expect that there would be more studies on hub genes than on non-hub genes due to their functional importance. Since we did not use information from the literature to obtain the functional association relationships between genes, it is interesting to examine the correlation between the gene connectivity and Pubmed hits. We implemented a script to search Pubmed with 'retina and gene name' for each gene in the network. We defined the number of returned papers as the Pubmed hit number. Larger hit number indicates more retina-related studies on the gene. Because some genes may have other aliases, we searched the Pubmed for official symbol and each alias and then merged the Pubmed IDs into a non-redundant Pubmed hit set. The median Pubmed hit numbers for non-hub and hub genes are 9 and 17, respectively (Supplementary Fig. S2), confirming that more studies were conducted on hub genes than on non-hub genes ($P < 0.02$, two sample *t*-test). We found that some extensively studied hub genes are those with the highest degrees (>90). Such examples include *Crx* (degree = 42, Pubmed hit number = 176), *Nr2e3* (33, 98), *Gnat1* (75, 956), *Abca4* (54, 164), *Pax6* (48, 334),*Rcvrn* (48, 380), *Sag* (47, 338), *Ctnnb1* (39, 91),*Gabrr1* (37, 163), *Guca1a* (36, 95), *Pde6b* (34, 122), *Shh* (33, 95) and *Rho* (31, 615) (Table 2). This result suggests that researchers have already captured and studied many important genes in the network. However, some hub genes have low Pubmed hit numbers (<5), suggesting they are likely top-candidate genes for further retina research. Such examples include *Accn1*, *Scn1a*, *Cacnb2*, *Ppef2*, *Ankrd33*, *Gngt1*, *Cacna2d3*, *Slc24a1*, *Atp1a3* and *Kcnj14*.

3.6 Novel function prediction

We are also able to predict gene functions based on network topology, and this method can complement the more traditional sequence homology approach. If the neighbors of a gene are enriched in a certain function, we predict that the gene itself also has that function (see Section 2). The power of such 'guilt by association' approach has already been shown in the cluster and hub gene analysis. Here we would like to make a systematic prediction of gene function in the retinal network. For each gene, we checked whether its neighbor genes are enriched for certain function and assigned the enriched function to the gene. Using this approach we predicted that 2875 GO function terms are associated with 315 genes ($P < 0.01$, Bonferroni correction, see Section 2). Among them, 1864 GO terms (64.8%) are known to be associated to the genes. In addition, 1011 novel functions are predicted for 200 genes (Supplementary Table S3).

We performed a web search to check whether some of our predictions are supported by other evidence. One example is *Angpt2* (angiopoietin 2). The protein encoded by this gene is an antagonist of *Angpt1*, and its deficiency can decelerate age-dependent vascular changes in the mouse retina (Feng *et al*., 2008). Our prediction suggests this gene have cytokine activity (GO:0005125). As a support, two recent publications indicate that *Angpt2* has cytokine activity and may induce endothelial cell apoptosis (Niedzwiecki *et al*., 2006; Peters, *et al*., 2007).

Another example is *Drd4*. *Drd4* (Dopamine receptor 4) is a G protein-coupled receptor, and its mutation have been associated with various behavioral phenotypes, including attention deficit/hyperactivity disorder and schizophrenia. Our prediction indicates that *Drd4* is related to photoreceptor activity. As support, a recent published paper indicates that *Drd4* regulates Adcy1 mRNA level and the cyclic AMP synthesis in photoreceptors, thus influences photoreceptor activity (Jackson *et al*., 2009).

Here we only showed a few examples to demonstrate that independent published work supported our gene function predictions. It is important to note that our results provide many interesting hypothesis for further experimental validation in the retinal research community.

3.7 Disease-associated gene

As one more application of this retinal gene network, we can predict disease-associated genes based on the assumption that a gene densely connected to known disease-associated genes also tends to be associated with the same or similar disease. If one gene's neighbors are enriched in genes that are known to be associated with a retinal disease, we predict that the gene is also associated with the disease. We used the annotation in the database RetNet (www.sph.uth.tmc.edu/RetNet/) for the definition of retinal disease genes.

Retinitis Pigmentosa (RP) is a hereditary and generally blinding disease that causes rod photoreceptors in the retina to gradually degenerate. According to the annotation in RetNet, 22 genes in the network are known to be RP-related genes (Supplementary Table S2). To evaluate the prediction power of the network analysis, we randomly selected 17 genes as training set and predicted whether other genes in the network (including the remaining five RP genes) are RP-associated genes. The genes were ranked by the *P*-value. This procedure was repeated 100 times (see Section 2). Each gene was assigned an average ranking. We then calculated the percentages of the 22 known RP-associated genes recovered given a ranking threshold (Fig. 3a). It is clear that the known disease genes are not homogenously distributed in the gene list. They tend to be at the top positions, suggesting that genes involved in the same or similar diseases are likely to be connected in the retinal gene network. We not only recovered some known disease genes, we were also able to

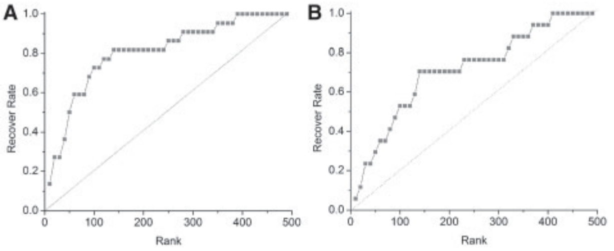

Fig. 3. Recovery rate of disease genes. (**A**) Retinitis pigmentosa (RP). (**B**) Cone-rod dystrophy (CRD).

predict some novel RP-associated genes (Supplementary Table S2). Some of the predicted genes are supported by other evidence. For example, *Pdc* has been considered as a potential candidate gene for RP in OMIM database. *Gnat1* is a transducin gene of visual impulse that performs the coupling between rhodopsin and cGMP-phosphodiesterase. Mutation in *Gnat1* has been reported to lead to a RP-related diseases, the Nougaret form of congenital stationary night blindness (Dryja *et al*., 1996). *Unc119* is a stimulus-response-related gene. Although it has not been reported to be mutated in any known form of retinal disease, its inactivation can cause photoreceptor dysfunction and slow retinal degeneration (Ishiba *et al*., 2007).

We performed the same analysis on cone-rod dystrophy (CRD), which is an inherited progressive disease that causes degeneration of cone and rod photoreceptor cells and often results in blindness. Based on RetNet annotation, 17 genes in the network are known to be CRD associated genes (Supplementary Table S4). Again, the known CRD-associated genes are more likely to be on top of the gene list. In addition, we also predicted some novel CRD-associated genes. Of them, *Ddb1* has been reported to be associated with a CRD related disease, macular dystrophy (Stohr *et al*., 1998).

3.8 MoReNet, a web-based interface

A web-based application, Mouse Retinal Network (MoReNet, http://bioinfo.wilmer.jhu.edu/morenet/) has been developed that allows the user to search for gene sets functionally related to a user-supplied gene list (Fig. 4). Different probability cutoffs can be selected to define the functional related gene pairs, ranging from 0.1 to $1e-5$. The search result is shown in interactive SVG figures to allow the user to zoom in or out and change parameters, such as node color, text color, text size, edge color, edge width. Nodes in the figure are placed by a Force-directed placement algorithm (Fruchterman and Rheingold, 1991), and can be manually rearranged to get a better layout.

4 DISCUSSION AND CONCLUSIONS

In this study, we integrated a variety of heterogeneous large-scale genomic datasets to construct a tissue-specific gene network that describes interactions between retinal genes in the mouse. Development of such networks has been made possible by the recent explosion of available high-throughput biological data, and the ongoing exponential growth of genome-wide data should make possible the continuing construction and refinement of such network models.

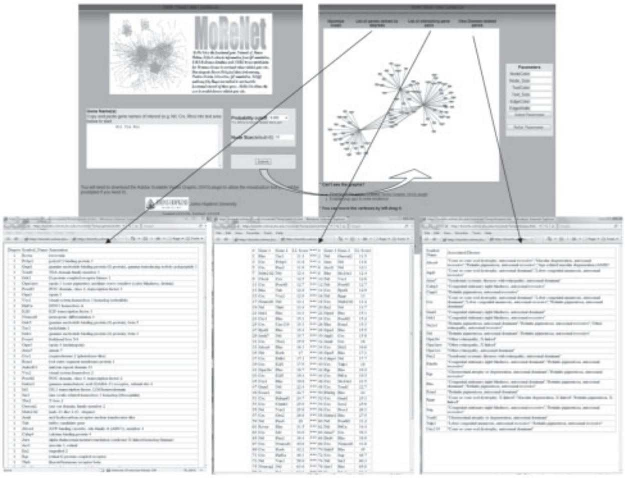

Fig. 4. Screenshot showing the interrelation of different WebPages. The user can search in the website to find gene sets functionally related to genes of interest.

Gene network analysis is an emerging approach to study biology. Previously, prediction of gene function relied heavily on homology comparison of protein structure or sequence. If a gene of unknown function does not have high sequence similarity to other genes, it is difficult to predict its function, and studies of such unknown genes have often relied upon genetic gain-of-function and loss-of-function experiments. Prediction based on sequence similarity only utilizes limited information. Gene network provides a new way to predict gene function based on the network topology. A similar concept also applies to disease gene prediction. Network analysis has been proven to be a powerful method to predict disease genes (Franke *et al*., 2006; Goh *et al*., 2007). Another merit of this method is that it can also predict the importance of genes based on their degrees in the network, thus experimental biologist can first study the vital hub genes. This can greatly accelerate the complete functional annotation of the genome.

For mammalian systems, universal gene networks have been constructed and analyzed. However, we believe that tissue-specific (or cell type-specific) networks are vital to understand tissue specificity given the fact that each cell has identical genomic DNA sequences. For this purpose, we selected a set of retina-related genes for construction of the network. We believe that any gene expressed in the retina may play a role in retinal function to some extent. However, considering our aim to construct a retina-specific gene functional network, we are interested in the genes that are 'specific' to retina and the relationship between them. Here, we use relative gene expression level to define retina-specific genes in addition to the known retina-related genes in literatures. If we use absolute gene expression level as criteria, many of genes in the network will be the house-keeping genes. The network would be essentially the same as a 'universal' gene network instead of a tissue specific gene network. The role of these universal genes is the maintenance of basic cellular functions, and not the specific function of the tissue of interest. Our study provides a careful analysis of the retinal gene network and facilitates more discoveries for retinal function studies in this community. To more fully understand a complex biological system such as the retina, the next step would be the analysis of the dynamics of the gene network, adding one more dimension (time or condition) to the system.

Our established computational methods can be readily applied to other tissues and used to create different tissue-specific gene networks. These gene networks will increase our understanding of functional relationships of genes in a tissue-specific manner. Due to the striking similarity between the mouse and human genomes, we expect it to be possible to extrapolate the predicted functional association of genes in mouse retina to that of human retina genes. Thus, our study can also help to predict genes related to human retinal diseases, and provide clues and possible targets of treatment.

Funding: National Institutes of Health (grant EY017589); Funds supporting the Guerrieri Center for Genetic Engineering and Molecular Ophthalmology, generous gift from Mr and Mrs Robert and Clarice Smith.

Conflict of Interest: none declared.

REFERENCES

Alon,T. *et al.* (1995) Vascular endothelial growth factor acts as a survival factor for newly formed retinal vessels and has implications for retinopathy of prematurity. *Nat. Med.*, **1**, 1024–1028.

Bateman,A. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.

Bertone,P. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.

Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Blanchette,M. *et al.* (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.*, **16**, 656–668.

Cawley,S. *et al.* (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.

Chen,S. *et al.* (1997) Crx, a novel Otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. *Neuron*, **19**, 1017–1030.

Chen,C.K. *et al.* (1999) Abnormal photoresponses and light-induced apoptosis in rods lacking rhodopsin kinase. *Proc. Natl Acad. Sci. USA*, **96**, 3718–3722.

Chen,J. *et al.* (2005) The rod photoreceptor-specific nuclear receptor Nr2e3 represses transcription of multiple cone-specific genes. *J. Neurosci.*, **25**, 118–129.

Delarasse,C. *et al.* (2009) Neural progenitor cell death is induced by extracellular ATP via ligation of P2X7 receptor. *J .Neurochem.*, **109**, 846–857.

Deng,M. *et al.* (2004) Mapping Gene Ontology to proteins based on protein-protein interaction data. *Bioinformatics*, **20**, 895–902.

D'Haeseleer,P. *et al.* (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, **16**, 707–726.

Dryja,T.P. *et al.* (1996) Missense mutation in the gene encoding the alpha subunit of rod transducin in the Nougaret form of congenital stationary night blindness. *Nat. Genet.*, **13**, 358–360.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Feng,Y. *et al.* (2008) Angiopoietin-2 deficiency decelerates age-dependent vascular changes in the mouse retina. *Cell Physiol. Biochem.*, **21**, 129–136.

Franke,L. *et al.* (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.

Fruchterman,T.M. and Rheingold,E.M. (1991) Graph drawing by force-directed placement. *Software-Pract. Exp.*, **21**, 1129–1164.

Goh,K.I. *et al.* (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Guan,Y. *et al.* (2008) A genomewide functional network for the laboratory mouse. *PLoS Comput. Biol.*, **4**, e1000165.

Haeseleer,F. *et al.* (2004) Essential role of Ca2+-binding protein 4, a Cav1.4 channel regulator, in photoreceptor synaptic function. *Nat. Neurosci.*, **7**, 1079–1087.

Harrison,P.M. *et al.* (2005) Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res.*, **33**, 2374–2383.

Hsiau,T.H. *et al.* (2007) The cis-regulatory logic of the mammalian photoreceptor transcriptional network. *PLoS One*, **2**, e643.

Ideker,T. and Sharan,R. (2008) Protein networks in disease. *Genome Res.*, **18**, 644–652.

Ishiba,Y. *et al.* (2007) Targeted inactivation of synaptic HRG4 (UNC119) causes dysfunction in the distal photoreceptor and slow retinal degeneration, revealing a new function. *Exp. Eye Res.*, **84**, 473–485.

Jackson,C.R. *et al.* (2009) Essential roles of dopamine D4 receptors and the type 1 adenylyl cyclase in photic control of cyclic AMP in photoreceptor cells. *J. Neurochem.*, **109**, 148–157.

Jansen,R. *et al.* (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.

Jeong,H. *et al.* (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Kim,T.H. *et al.* (2005) A high-resolution map of active promoters in the human genome. *Nature*, **436**, 876–880.

Kim,W.K. *et al.* (2008) Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy. *Genome Biol.*, **9** (Suppl. 1), S5.

Lage,K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Lee,H.K. *et al.* (2004a) Coexpression analysis of human genes across many microarray data sets. *Genome Res.*, **14**, 1085–1094.

Lee,I. *et al.* (2004b) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.

Letovsky,S. and Kasif,S. (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, **19** (Suppl. 1), i197–i204.

Li,J. *et al.* (2009) P2X7 nucleotide receptor plays an important role in callus remodeling during fracture repair. *Calcif. Tissue Int.*, **84**, 405–412.

Liu,X. *et al.* (2004) AIPL1, the protein that is defective in Leber congenital amaurosis, is essential for the biosynthesis of retinal rod cGMP phosphodiesterase. *Proc. Natl Acad. Sci. USA*, **101**, 13903–13908.

Lord,P.W. *et al.* (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

Lu,L.J. *et al.* (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.*, **15**, 945–953.

Mears,A.J. *et al.* (2001) Nrl is required for rod photoreceptor development. *Nat. Genet.*, **29**, 447–452.

Mingam,R. *et al.* (2008) In vitro and in vivo evidence for a role of the P2X7 receptor in the release of IL-1 beta in the murine brain. *Brain Behav. Immun.*, **22**, 234–244.

Niedzwiecki,S. *et al.* (2006) Angiopoietin 1 (Ang-1), angiopoietin 2 (Ang-2) and Tie-2 (a receptor tyrosine kinase) concentrations in peripheral blood of patients with thyroid cancers. *Cytokine*, **36**, 291–295.

Niehrs,C. and Pollet,N. (1999) Synexpression groups in eukaryotes. *Nature*, **402**, 483–487.

Oti,M. *et al.* (2006) Predicting disease genes using protein-protein interactions. *J. Med. Genet.*, **43**, 691–698.

Pena-Castillo,L. *et al.* (2008) A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. *Genome Biol.*, **9** (Suppl. 1), S2.

Peters,S. *et al.* (2007) Angiopoietin modulation of vascular endothelial growth factor: effects on retinal endothelial cell permeability. *Cytokine*, **40**, 144–150.

Pujana,M.A. *et al.* (2007) Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat. Genet.*, **39**, 1338–1349.

Qian,J. *et al.* (2005) Identification of regulatory targets of tissue-specific transcription factors: application to retina-specific gene regulation. *Nucleic Acids Res.*, **33**, 3479–3491.

Reich,M. *et al.* (2004) GeneCluster 2.0: an advanced toolset for bioarray analysis. *Bioinformatics*, **20**, 1797–1798.

Reim,K. *et al.* (2009) Aberrant function and structure of retinal ribbon synapses in the absence of complexin 3 and complexin 4. *J. Cell Sci.*, **122**, 1352–1361.

Sanuki,R. *et al.* (2010) Panky, a novel photoreceptor-specific ankyrin repeat protein, is a transcriptional cofactor that suppresses CRX-regulated photoreceptor genes. *FEBS Lett.*, **584**, 753–758.

Scherer,S.W. *et al.* (1996) Gene structure and chromosome localization to 7q21.3 of the human rod photoreceptor transducin gamma-subunit gene (GNGT1). *Genomics*, **35**, 241–243.

Sharon,D. *et al.* (2002) Mutated alleles of the rod and cone Na-Ca+K-exchanger genes in patients with retinal diseases. *Invest. Ophthalmol. Vis. Sci.*, **43**, 1971–1979.

Stohr,H. *et al.* (1998) Refined mapping of the gene encoding the p127 kDa UV-damaged DNA-binding protein (DDB1) within 11q12-q13.1 and its exclusion in Best's vitelliform macular dystrophy. *Eur. J. Hum. Genet.*, **6**, 400–405.

Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.

Wang,Y. *et al*. (2009) In vitro study of the effects of lens extract on rat retinal neuron survival and neurite outgrowth. *Ophthalmic Res.*, **42**, 29–35.

Whisstock,J.C. and Lesk,A.M. (2003) Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.*, **36**, 307–340.

Wu,N. *et al*. (2009) alpha-Crystallin downregulates the expression of TNF-alpha and iNOS by activated rat retinal microglia in vitro and in vivo. *Ophthalmic Res.*, **42**, 21–28.

Xi,J. *et al*. (2003) A comprehensive analysis of the expression of crystallins in mouse retina. *Mol. Vis.*, **9**, 410–419.

Yu,X. *et al*. (2006) Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.*, **34**, 4925–4936.

Zeitz,C. *et al*. (2006) Mutations in CABP4, the gene encoding the Ca2+-binding protein 4, cause autosomal recessive night blindness. *Am. J. Hum. Genet.*, **79**, 657–667.

Zhang,W. *et al*. (2004) The functional landscape of mouse gene expression. *J. Biol.*, **3**, 21.