2009

Using a seed-network to query multiple large-scale gene expression datasets from the developing retina in order to identify and prioritize experimental targets

Timothy Alcon
Iowa State University

Using a seed-network to query multiple large-scale gene expression datasets from the developing retina in order to identify and prioritize experimental targets

by

Timothy C. Alcon

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Vasant G. Honavar, Co-major Professor
M. Heather West Greenlee, Co-major Professor
Jeanne M. Serb

Iowa State University

Ames, Iowa

2009

DEDICATION

I would like to dedicate this thesis to all the family and friends who helped me remain relatively sane during my graduate studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank those who helped me with various aspects of conducting research and the writing of this thesis: Dr. Vasant Honavar and Dr. Heather Greenlee for their guidance, patience and support throughout this research and the writing of this thesis; Laura Hecker for being a great collaborator; Fadi Towfic for many helpful discussions; Dr. Jeanne Serb for serving on my committee; Kathy Weiderin, Trish Stauble, Linda Dutton and Lanette Woodard for much helpful advice and assistance along the way.

GENERAL INTRODUCTION

Introduction and Literature Review

As more and more biological data is accumulated and stored in databases, we are faced with the question of how all this data can be effectively used to advance our understanding of how biological systems work. Are there ways of combining multiple datasets from different experiments and different labs that can tell us more than each of the datasets reveals individually? There are many different types of high-throughput data (eg. gene expression data, transcription factor binding data, yeast two-hybrid data, Gene Ontology data, etc.), each of which look at different parts of the biological puzzle. Gene expression data consists of measurements of how much mRNA or protein is being produced from the corresponding genes, but these measurements are often noisy (Claverie (1999)). Transcription factor binding data consists of potential binding sites for transcription factors (TFs), usually found by computationally searching the genome for matches to a known binding motif for a TF, but many of these matches aren't regulated by that TF *in vivo* since this method only takes into account the motif sequence and not whether biological conditions, such as subcellular location and timing of expression, actually permit the TF and its hypothetical target to interact. Yeast two-hybrid (Y2H) data consists of pairs of proteins that were found to bind to each other when expressed in genetically engineered yeast, but Y2H data are known for having large numbers of false positives (Serebriiskii et al. (2000)). Gene Ontology data consists of gene annotations from a controlled vocabulary of terms that define a gene's molecular function, biological process and cellular component, but some genes are less thoroughly annotated than others, and some annotations are based on less reliable evidence. Because each type of data provides a different

kind of biological information and each type has its own strengths and weaknesses, it would seem beneficial if they could be combined in such a way that they could complement and reinforce each other, and there have been some recent attempts to do so (Gunsalus et al. (2005); Rhodes et al. (2005); Xia et al. (2006); Pujana et al. (2007); English and Butte (2007)). Pujana et al. (2007), for example, combined gene expression data, phenotypic similarity data, genetic interaction data and protein physical interaction data in order to find potential functional associations with cancer genes/proteins. In the work that follows, however, we restrict ourselves specifically to using gene expression datasets that were generated using different platforms in different labs, itself a non-trivial problem. We first propose a way to combine five time-series gene expression datasets (Akimoto et al. (2006); Blackshaw et al. (2004); Dorrell et al. (2004); Liu et al. (2006); Zhang et al. (2006)) to create a composite network of robust gene correlations. It has long been thought that if genes are correlated across multiple conditions or time points, it may be evidence that they are co-regulated, and studies have found support for this idea (Wolfe et al. (2005); Rhodes et al. (2005)). Next, we propose a method for starting with a seed network of genes and using it to query the composite correlational network in a way that permits us to rank other genes for possible inclusion in an extended seed network.

The biological system explored in this work is the differentiation of retinal progenitor cells into rod photoreceptors in the murine retina. During development, retinal progenitor cells differentiate into six major retinal cell types, of which photoreceptors are one. Photoreceptors are the cells that detect incoming light and are comprised of the rod photoreceptors, which can detect lower levels of light, but are restricted to black and white vision, and the cone photoreceptors, which are able to distinguish color. Knowing more about this particular process has great potential theraputic value. Retinal diseases such as retinitis pigmentosa and macular degeneration cause loss of vision due to the death of photoreceptors. Some model organisms, such as frog and zebrafish, have the ability to replace damaged photoreceptors (Adler and Raymond (2008)), but in the mammalian retina these cells do not regenerate naturally. It has been proposed that transplanting stem or progenitor cells into the damaged retina might effect some degree of repair (Chacko et al. (2000); Sakaguchi et al. (2003, 2004); Hoffelen et al.

(2003)), but MacLaren et al. (2006) have shown that post-mitotic rod precursors show much greater ability to integrate into the existing retinal circuitry than are less differentiated cells. This being the case, it seems that if we knew more about the network of gene interactions that influences rod photoreceptor differentiation, it might be possible to bias retinal progenitor cells toward a rod photoreceptor cell fate, enhancing their potential for repairing retinal damage.

The two primary challenges faced in the research presented here were figuring out effective approaches for: 1) integrating gene expression data across disparate platforms; and 2) given a correlational network, extending a seed network by adding to it genes likely to be related to the seed genes. In the following, I will briefly survey published work related to these two challenges.

Combining cross-platform gene expression datasets

There are huge number of publicly available gene expression datasets (more than 350,000 GSM samples alone in GEO, NCBI's gene expression repository), and this number is increasing rapidly. It would be of enormous benefit to researchers to have a simple effective procedure for comparison and integration of multiple datasets relevant to a particular topic of interest. This is complicated by the fact that many different types of platform are involved. There are different technologies used (e.g. microarrays and SAGE), different methods of implementing a given technology (e.g. two-color microarray and oligonucleotide microarray), and different designs within a particular method (e.g. different versions of Affymetrix microarrays, even for the same species). Methods for integrating gene expression datasets across platforms fall into two major categories: low-level methods, where the actual (transformed) expression measurements are combined (Stevens and Doerge (2005); Warnat et al. (2005); Choi et al. (2003); Marot et al. (2009); Hong et al. (2006)); and high-level methods, where the datasets are each analyzed separately and then those results are combined (Zhou et al. (2005); Rajaram (2009); Parmigiani et al. (2004); Conlon et al. (2006, 2007); Griffith et al. (2005); Lee et al. (2004)).

Low-level integration

Effect sizes are are often utilized in many types of meta-analysis. An effect size is a standardized unitless measure of the size of the effect of a treatment. This is different from a p-value which depends on both effect size and sample size and gives the probability that an observed difference is due to chance. There are several different effect size estimates that can be used. The standardized means difference was used as the effect size estimate by Choi et al. (2003), who advocate using a homogeneity test to determine whether a fixed effects model or random effects model would be more appropriate for the datasets that are to be combined. They also apply a Bayesian approach to combine effect sizes across studies and conclude that with appropriate prior information it can provide a more flexible and robust strategy. Stevens and Doerge (2005) use the signal log ratio (SLR) as an effect-size estimate. In a study with simulated datasets, they found that the meta-analysis SLR estimates were closer to the true SLR values than the SLR estimates from the individual simulated studies. A moderated effect-size method is proposed by Marot et al. (2009), who compare it to standard effect sizes as well as to combination of p-values, concluding that although the moderated effect size was an improvement over standard effect sizes, the combination p-value approach resulted in greater sensitivity. One difficulty with using effect size for a meta-analysis of gene expression data, however, is that it makes the assumption that the data are normally distributed. This is arguably true for microarray data, however it has been suggested that SAGE data follow a Poisson distribution (Cai et al. (2004)).

Two other studies use low-level approaches that do not depend on effect size. Warnat et al. (2005) compared two different methods: median rank scores, where datasets are transformed to similar numerical ranges by replacing expression values in one study with those from another study based on the relative ranks of the expression values; and quantile discretization, where the expression values for each study are discretized into a set number of bins, using quantiles of the expression values as cut points. Both methods were found to improve accuracy of an SVM classifier over performance on the non-integrated datasets, but for most datasets there was no significant difference between the two methods. Hong et al. (2006) adapt the rank

product method from Breitling et al. (2004) so that it may be used for combining microarray studies from different platforms. The method consists of: 1) computing pairwise ratios between treatment and control arrays; 2) ranking these ratios within each comparison; 3) computing the rank product by taking the product of ratios at the same rank across comparisons and then taking the k^{th} root; 4) independently permuting expression values within each array, then repeating steps 1-3; and 5) repeating step 4 some number of times to form a reference distribution and determining the p-value and FDR for each gene.

High-level integration

High-level approaches to combining gene expression datasets are not concerned with finding a lab and platform independent representation of expression levels. Instead they only compare expression levels within datasets and then look for any way of combining this higher-level information across datasets in such a way as could cast light on the underlying biological processes. Since the goal is less narrowly specified than for low-level integration, methods for high-level integration are more varied. Parmigiani et al. (2004) fit for each study and gene a Cox proportional hazards model, with gene expression first divided by its standard deviation to improve the comparability across platforms of the resulting coefficients. Zhou et al. (2005) suggest the use of 2nd-order correlation for integrating data across platforms. They first compute all pairwise gene correlations within each dataset (using jackknife Pearson correlation), defining as doublets those pairs that meet the correlation threshold in enough datasets. They then take the correlation between the vectors of first-order expression correlations of non-overlapping doublets, defining those that meet the threshold as quadruplets, which they argue are likely to be functionally linked. A hierarchical Bayesian model is used by Conlon et al. (2006) to combine probabilities of differential expression across multiple studies (but from the same platform, cDNA microarray), providing the gene-specific posterior probability of differential expression and Bayesian estimates of false discovery rates. This method was compared with the method from Choi et al. (2003) by Conlon et al. (2007), who found in simulations that combining probabilities (high-level) outperformed combining standardized gene

expression measures (low-level). Rajaram (2009) attempts to identify internal consistency sets comprised of groups of genes within which pairwise correlations remain, not necessarily high, but consistent across datasets. He begins with an ICS consisting of a random set of N genes and then ranks all genes according to their consistency across datasets with the genes in the current ICS. The N top-ranked genes become the new ICS. These steps are repeated until a fixed point is reached. The whole process can the be repeated to explore the space of ICSs of size N. Griffith et al. (2005) take the Pearson correlation for all gene pairs within each platform and then average those correlations across datasets, keeping those above a particular threshold. Lee et al. (2004) also take the Pearson correlation for all gene pairs within each platform, but rather than averaging correlations across datasets and then seeing whether they then meet a threshold, they use a vote counting method, keeping those gene correlations that met a threshold in at least three datasets. This seems advantageous over the averaging method used by Griffith et al. (2005), since it prevents outliers from shifting the median expression above or below the threshold. The method used in our work is most similar to that of Lee et al. (2004). We found their approach to be highly suitable for use in this study, since it is conceptually simple, yet was shown to be effective in terms of finding links between genes that were confirmed by a semantic similarity metric based on the overlap of Gene Ontology annotations for each pair of linked genes. In our work, however, we used the Spearman rank correlation instead of the Pearson correlation, since the Pearson correlation assumes a normal distribution and, as mentioned above, it has been suggested that SAGE data follow a Poisson distribution Cai et al. (2004). And since we were working with five datasets (rather than 60) we retained gene correlations that were supported by two or more datasets. This is also, to our knowledge, the first study to combine time-series gene expression data across platforms.

Extending a seed network within a larger given network

It's often the case that a researcher will know a set of genes related to a particular biological process of interest and would like to determine what other genes are most likely to also be related to that process. We restrict ourselves here to those cases where a gene network

(usually very large) is either publically available or can be constructed from available data. The links between genes may be based on traditional biological approaches, such as chromatin immunoprecipitation, but for large networks will usually be constructed from high-throughput data (e.g. microarray or yeast two-hybrid). In this case, the links may not be reliable and will often have weights on the edges giving some summary measure, such as correlation, that helps to indicate the likelihood of that link being real. The problem is then, given a large weighted gene network and some small subnetwork within it of genes that are functionally related to each other, what is the best way to expand that subnetwork by determining which neighboring genes should be included?

Bader (2003) uses a greedy algorithm isomorphic to a single-source shortest path search, with the queue initialized to have multiple entries at distance 0, if the length of an edge is identified with the negative logarithm of the weight on that edge. Cabusora et al. (2005) extract a sub-network that is spanned by the seed nodes by including genes that appear on short paths between them, which are determined using Dijkstra's and Yen's algorithms. A Monte Carlo method is used by Asthana et al. (2004) to sample many binary networks from a weighted graph where the weights are treated as probabilities. The fraction of the sample with a path to the seed genes is the estimated probability used to rank a gene. Using a random walk with restart from a set of seed genes, with the walker's probability of traversing an edge based on the weights of all adjacent edges was proposed by Can et al. (2005). The percentage of time the walker spends at a given gene is the estimated probability used to rank that gene. This is more computationally efficient than the Monte Carlo simulation technique of Asthana et al. (2004). Köhler et al. (2008) compare using random walk with restart with using a diffusion kernel method, which can be thought of as a different kind of random walk based on matrix exponentiation. They found that random walk with restart had better performance than the diffusion kernel method. Hashimoto et al. (2004) use an approach that, in a directed probabilistic Boolean network, adds genes that enhance the collective strength of connections within the seed network based on the coefficient of determination (Dougherty et al. (2000)) and the Boolean-function influence (Shmulevich et al. (2002)). Li and Horvath (2007) generalize to

multiple nodes the Topological Overlap Matrix (Ravasz et al. (2002)), in which the topological overlap of two nodes gives their similarity as based on the commonality of the nodes they each connect to. At each iteration the gene associated with the highest MTOM value is added to the seed network.

The method used in this work is 1) for each seed gene to produce a list of genes ranked according to both the average strength of correlation with the seed gene and the number of datasets in which the correlation met the threshold; and 2) to produce a list of genes that show high correlation with multiple seed genes. For both lists, each gene is annotated with information from KEGG and the Gene Ontology. A biologist may then sort the lists by any of the mentioned attributes as desired in order to evaluate the suitability of adding a particular gene to the seed network.

Thesis Organization

Inclusion of journal article

The central chapter of this thesis is modified from a journal article published in 2008 by the Journal of Bioinformatics and Biology Insights, and on which Timothy C. Alcon was a co-first author.

Corrections

A couple of minor corrections have been introduced in this reprint of the journal article for inconsistencies discovered subsequent to the article's publication. Specifically, Table 2 and Figure 2, together with the relevant text, were amended because the correlations used to generate them were incorrect for two of the datasets. All other figures and tables, including the supplementary tables, were generated through independent means and do not require any alteration from their published form. Furthermore, the minor corrections that have been introduced here do not substantially affect any of the arguments or conclusions set forth in the article.

Author responsibilities

Timothy C. Alcon carried out all of the statistical and computational analyses and created the tables, figures and text associated with those analyses.

Laura A. Hecker constructed the network from the literature and the expanded network and created the figures and text associated with those analyses.

Drs. Greenlee and Honavar conceived the study, designed the experiments, and participated in writing and editing the manuscript.

USING A SEED-NETWORK TO QUERY MULTIPLE LARGE-SCALE GENE EXPRESSION DATASETS FROM THE DEVELOPING RETINA IN ORDER TO IDENTIFY AND PRIORITIZE EXPERIMENTAL TARGETS

A paper modified from an article published in the Journal of Bioinformatics and Biology Insights[1]

Laura A. Hecker[2,3], Timothy C. Alcon[2,4], Vasant G. Honavar[5] and M. Heather Greenlee[6]

Abstract

Understanding the gene networks that orchestrate the differentiation of retinal progenitors into photoreceptors in the developing retina is important not only due to its therapeutic applications in treating retinal degeneration but also because the developing retina provides an excellent model for studying CNS development. Although several studies have profiled changes in gene expression during normal retinal development, these studies offer at best only a starting point for functional studies focused on a smaller subset of genes. The large number of genes profiled at comparatively few time points makes it extremely difficult to reliably infer

[1]Journal of Bioinformatics and Biology Insights, 2008; 2: 401412.

[2]Laura A. Hecker and Timothy C. Alcon are joint first authors of this work

[3]Interdepartmental Neuroscience Program, Iowa State University, Ames, IA 50011

[4]Bioinformatics and Computational Biology Graduate Program, Iowa State University, Ames, IA 50011

[5]Department of Computer Science, Bioinformatics and Computational Biology Graduate Program, Center for Computational Intelligence, Learning and Discovery, Iowa State University, Ames, IA 50011

[6]Department of Biomedical Sciences, Interdepartmental Neuroscience Program, Bioinformatics and Computational Biology Graduate Program, Center for Computational Intelligence, Learning and Discovery, Iowa State University, Ames, IA 50011

Correspondence: M. Heather West Greenlee, Department of Biomedical Sciences, Interdepartmental Neuroscience Program, Bioinformatics and Computational Biology Graduate Program, Center for Computational Intelligence, Learning and Discovery, Iowa State University, Ames, IA 50011. Email: mheather@iastate.edu

gene networks from a gene expression dataset. We describe a novel approach to identify and prioritize from multiple gene expression datasets, a small subset of the genes that are likely to be good candidates for further experimental investigation. We report progress on addressing this problem using a novel approach to querying multiple large-scale expression datasets using a seed network consisting of a small set of genes that are implicated by published studies in rod photoreceptor differentiation. We use the seed network to identify and sort a list of genes whose expression levels are highly correlated with those of multiple seed network genes in at least two of the five gene expression datasets. The fact that several of the genes in this list have been demonstrated, through experimental studies reported in the literature, to be important in rod photoreceptor function provides support for the utility of this approach in prioritizing experimental targets for further experimental investigation. Based on Gene Ontology and KEGG pathway annotations for the list of genes obtained in the context of other information available in the literature, we identified seven genes or groups of genes for possible inclusion in the gene network involved in differentiation of retinal progenitor cells into rod photoreceptors. Our approach to querying multiple gene expression datasets using a seed network constructed from known interactions between specific genes of interest provides a promising strategy for focusing hypothesis-driven experiments using large-scale omics data.

Introduction

Blinding degenerative retinal diseases including retinitis pigmentosa and macular degeneration are characterized by a loss of photoreceptors. At present there is no way to replace retinal cells lost due to disease or injury because differentiated retinal cells are unable to regenerate. Various stem and/or progenitor cell populations have been proposed as a potential source of transplantable cells to replace lost cells in the damaged retina. The retina is composed of five major neuronal types and one glial cell type that all originate from the same pool of progenitor cells. The rod photoreceptors, the most numerous among retinal cells, together with cone photoreceptors, are responsible for transduction of light and are required for vision. Recent studies demonstrate that post-mitotic rod precursors are able to differentiate and fully integrate into

the damaged retina, whereas less differentiated cells are not (MacLaren et al. (2006)). Understanding the network of genes that orchestrate the differentiation of retinal progenitors may make it possible to bias expanded stem cell populations to generate rod precursors.

Large-scale gene expression profiling is aimed at helping to understand how genes influence each other in networks, which then control cell fate commitment and differentiation. There are a number of published studies that have profiled changes in gene expression during normal retinal development (Blackshaw et al. (2001, 2004); Diaz et al. (2003); Dorrell et al. (2004); Yu et al. (2003)). However, the large number of genes profiled at comparatively few time points or conditions presents significant statistical challenges in inference of genetic networks from any given dataset. One way to more effectively understand relationships between genes is to increase the number of expression measurements for a given gene, and/or focus the investigation on a small number of genes of interest (or between clusters of genes that have similar expression profiles) (Zhou and Mao (2006)). Approaches that leverage existing biological knowledge (e.g. experimentally determined interactions among a small set of genes) to focus the analysis of data from large-scale gene expression studies are beginning to be explored (Bader (2003); Cabusora et al. (2005); Can et al. (2005); Dougherty et al. (2000); Hashimoto et al. (2004); Shmulevich et al. (2002)). Of particular interest is the use of such approaches to prioritize targets for further investigation using traditional experimental techniques.

In this study, we explore an approach to integrated analysis of multiple gene expression datasets in the context of a set of experimentally established relationships between genes. We used the data from five previously published expression studies (Akimoto et al. (2006); Blackshaw et al. (2004); Dorrell et al. (2004); Liu et al. (2006); Zhang et al. (2006)) that have provided gene expression data for large numbers of genes under comparable conditions. We queried the resulting datasets using a seed network of genes known to play key roles during rod genesis and differentiation (Ahmad et al. (1998); Chen et al. (1997); Cheng et al. (2004); Furukawa et al. (2002, 1997); Green et al. (2003); Mears et al. (2001); Nishida et al. (2003); Pennesi et al. (2003); Rutherford et al. (2004); Zhang et al. (2004)). We hypothesize that additional genes important for rod genesis and differentiation are likely to be highly positively

or negatively correlated with genes that belong to the seed network. We generated a list of such candidate genes based on the correlation of their expression with genes in the seed network. To increase the robustness of analysis, we selected those genes that are correlated with multiple seed network genes in at least two of the five datasets. We further prioritized the resulting candidate genes, based on their gene ontology annotations, evidence of their membership in known cellular signaling pathways, and biological knowledge (whenever such knowledge is available). Using this approach, we identified genes whose expression levels are correlated with multiple genes of interest. Of these, 986 genes are positively correlated with multiple genes of interest and 531 are negatively correlated with multiple genes of interest. We short-listed 7 genes or groups of genes from the list of 986 candidates for inclusion in a hypothesized rod network that extends our seed network. We believe that our results demonstrate the utility of querying multiple large-scale gene expression profiles using a seed network to prioritize genes for further investigation using detailed experimental studies.

Materials and Methods

Datasets measuring gene or protein expression in the developing mouse retina

Datasets measuring gene or protein expression in the developing mouse retina at multiple time points include: SAGE (serial analysis of gene expression) of whole retina (Blackshaw et al. (2004)), two Affymetrix microarrays of whole retina using the Mu74Av2 chip (hereafter referred to as Mu74Av2_1 (Dorrell et al. (2004)) and Mu74Av2_2 (Liu et al. (2006)), one cDNA microarray of whole retina (Zhang et al. (2006)), one Affymetrix microarray of only developing rod progenitors using the MOE430.2.0 chip (Akimoto et al. (2006)), and 2D PAGE (polyacrylamide gel electrophoresis) of whole retina (Barnhill and Greenlee personal communication).

ID mapping

Genes or proteins from each of these datasets were matched by Entrez gene ID. These IDs were determined using NCBIs gene database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=gene) (Maglott et al. (2007)) and WebGestalt (http://bioinfo.vanderbilt

.edu/webgestalt/) (Zhang et al. (2005)). One difficulty with cross-platform studies is that each microarray probe or SAGE tag must be mapped to some common set of gene identifiers. It is very often the case that more than one probe or tag will be mapped to the same gene, with the possibility that the different probes or tags represent alternative splicings of the same gene. There are three possible approaches to this problem. One is to keep expression measurements for each probe or tag separate, as different versions of a gene. This fails to solve the problem since there is currently no good way to match equivalent splicings of the same gene across platforms. Another approach is to get rid of any genes with ambiguous mappings. This approach ends up throwing away a lot of potentially helpful data. The third possibility is to combine the expression measurements for probes or tags that map to the same gene. The drawback of this method is that if the different probes or tags represent valid alternative splicings of the same gene, then these different splicings may in fact have different biological roles and hence different patterns of expression. However it at least provides an approximate matching and avoids throwing away valuable data. In cases where multiple SAGE tags or 2D PAGE spots mapped to a single gene, we summed the tags/spots expressions to arrive at a total expression for the gene. In cases where multiple microarray probes mapped to a single gene, we took the median of the probes expressions to arrive at a total expression for the gene.

Gene and pathway annotation

KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways and GO (Gene Ontology) annotations were retrieved using WebGestalt (Zhang et al. (2005)). The most highly represented pathways in the table of correlations with multiple genes (supplementary data) were determined by grouping all genes containing a pathway annotation by the given annotation. Signaling pathways represented by five or more gene members were considered highly represented.

Results

Cross-dataset comparisons

In determining how well gene expression correlates across different gene expression datasets, it is not valid to directly compare expression values since different protocols and different normalization methods will result in wide variations in expression values even if the same microarray and biological conditions are used. Where different platforms are used, different pairs of datasets will also have different genes in common. Hence, we chose to use the correlation of correlations, or rc (Lee et al. (2003)) to assess the degree to which pairwise gene expression correlations compare across each pair of datasets. SAGE expression measurements likely follow a Poisson distribution (Cai et al. (2004)), though the often-used Pearson correlation assumes a normal distribution. Thus, we instead use a Spearman rank correlation version of the rc, which doesnt assume any particular distribution, but rather the relative ranks of the expression values (for example if expression values for a set a genes were 5.74, 2.18, 3.65 and 9.13, then their ranks relative to one another would be 3, 1, 2 and 4). The rc between each pair of datasets, computed using the R statistical software (http://www.r-project.org) (Ihaka and Gentleman (1996)), is given in Table 1. The most highly correlated pair of datasets had a correlation value of 0.33. Significance was computed in R by means of permutation testing, which yielded p-values < 0.001 for each pair of datasets except when one of them was the 2D PAGE data set, in which case the p-values ranged from 0.016 to 0.574. The relatively low degree of agreement between datasets is not especially surprising in light of published comparisons of mRNA gene expression data from multiple studies involving overlapping or even the same sets of genes (Haverty et al. (2004); Kuo et al. (2002); Tan et al. (2003)). These results suggest that inference of gene networks from individual gene expression datasets has to be approached with caution.

Seed network construction

Given the low degree of agreement among the different gene expression datasets, it is natural to question how feasible it is to infer gene networks from gene expression data. In order to

Table 1 Correlation of correlations values between each of the gene expression datasets. In calculating each correlation of correlations, only the subset of genes in common between the two datasets was used. This subset was different for each pair of datasets. SAGE = SAGE data from whole retina (Blackshaw et al. 2004); MOE430.2.0 = Affymetrix microarray data from developing rod progenitors (Akimoto et al. (2006)); Mu74Av2_1 = Affymetrix microarray data from whole retina (Dorrell et al. (2004)); Mu74Av2_2 = Affymetrix microarray data from whole retina (Liu et al. (2006)); cDNA microarray = cDNA microarray data from whole retina (Zhang et al. (2006)); 2DGE = 2D-PAGE data from whole retina (Barnhill and Greenlee, personal communication). * $p < 0.001$, ** $p < 0.02$, *** $p < 0.05$.

	SAGE	**MOE430.2.0**	**Mu74Av2_1**	**Mu74Av2_2**	**cDNA Microarray**	**2DGE**
SAGE		0.1*	0.23*	0.12*	0.09*	0.05***
MOE430.2.0	0.1*		0.18*	0.09*	0.04*	0
Mu74Av2_1	0.23*	0.18*		0.33*	0.09*	0.07**
Mu74Av2_2	0.12*	0.09*	0.33*		0.02*	0.06**
cDNA Microarray	0.09*	0.04*	0.09*	0.02*		0.06
2DGE	0.05***	0	0.07**	0.06**	0.06	

address this question, we used an experimentally verified network against which a network inferred from expression data could be validated. We relied on results of experimental studies of retinal development to identify a set of 10 genes that have been implicated in rod photoreceptor development to include in a seed network to serve as a basis for validation (Figure 1). The edges between genes in the network represent several types of links including non-directional interactions inferred from knockout studies (Green et al. (2003); Rutherford et al. (2004)) indirect effects on expression inferred from knockout studies (Zhang et al. (2004)), phosphorylation events inferred from mutation and transfection experiments (Weinberg (1995)), and direct transcriptional control of one gene by another (Ahmad et al. (1998); Chen et al. (1997); Cheng et al. (2004); Furukawa et al. (2002, 1997); Mears et al. (2001); Nishida et al. (2003); Pennesi et al. (2003)).

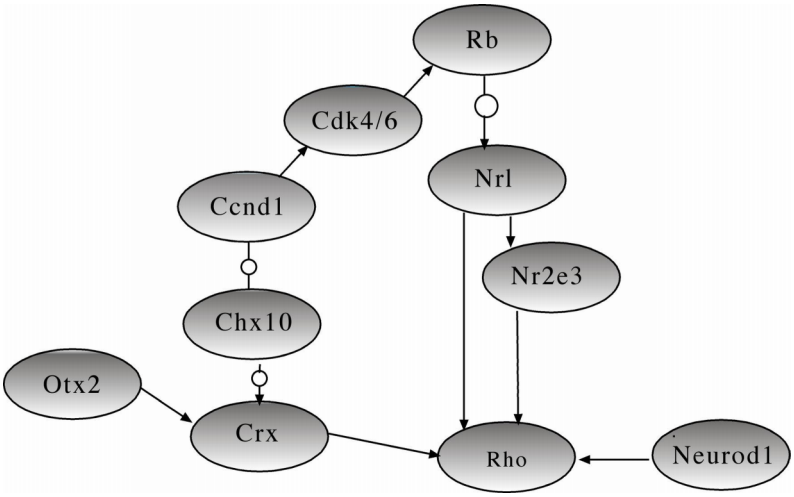

Figure 1 Representation of an intrinsic seed network controlling rod photoreceptor development. The network was constructed based on published experimental evidence and is made up of ten genes. Direct relationships between seed genes are indicated by arrows and indirect relationships are shown as arrows interrupted by circles.

Reconstruction of seed network from expression data

Having constructed a seed network to serve as a basis for testing the feasibility of inferring gene networks from gene expression data, we proceeded to explore whether the links between

the ten seed network genes (Figure 1) can in fact be reconstructed using one or more gene expression datasets (recall that the links between seed network genes reflect interactions between genes that are supported by published experimental studies).

We examined the pairwise correlations in expression between genes included in the seed network in each of the five mRNA expression datasets. The 2D gel electrophoresis (2DGE) dataset was omitted since none of the seed network genes were identified in it. In this analysis, a link between a pair of seed network genes is supported by a dataset if the corresponding genes were positively or negatively correlated in that dataset, with the absolute value of correlation greater than or equal to 0.65. Our choice of the threshold of 0.65 for correlation was influenced by similar choices in previous studies (Griffith et al. (2005); Gunsalus et al. (2005); Lee et al. (2004)) that have revealed biologically relevant links between coexpressed genes. Interestingly, no single dataset supported all six positive links in the seed network. One of the datasets supported five links, one dataset supported four links, two supported three links and one supported only one link (Table 2). We then proceeded to examine whether multiple datasets could be combined to reliably reconstruct the seed network from gene expression data. The resulting network (Figure 2) shows a link between a pair of seed network genes whenever the pairwise correlation between the expression levels of the corresponding genes is greater than or equal to +0.65 or less than or equal to -0.65 in at least 2 of the five datasets. Links depicting positive correlation are shown in blue and those depicting negative correlation are shown in red. Five of the six positive links in this reconstructed network (Figure 2) are also present as links in the original seed network (Table 2). In addition to the positive links there are four negative links based on the observed negative correlations between the seed network genes in the reconstructed network. Interestingly, the negative links partition the network into two sets of genes, one consisting of genes expressed by proliferating retinal progenitors (Chen and Cepko (2000); Sicinski et al. (1995); Zhang et al. (2004)) and the other consisting of genes expressed by cells in the process of differentiating into rod photoreceptors (Cheng et al. (2004); Furukawa et al. (2002); Mears et al. (2001); Morrow et al. (1998)). The success of this approach in recovering a majority of the links in the seed network, in spite of the relatively low

Table 2 Datasets supporting each positive edge between all pairs of genes shown to be linked in Figure 2. Datasets supporting a particular link between seed genes (based on correlation) are marked with an X. The last column indicates whether that edge was present in the network based on the literature (Figure 1).

	SAGE	**MOE430.2.0**	**Mu74Av2_1**	**Mu74Av2_2**	**cDNA Microarray**	**Original Seed Network**
Ccnd1-Cdk4		x	x	x	x	yes
Cdk4-Rb		x	x			yes
Crx-Nrl	x		x			no
Nrl-Nr2e3	x	x			x	yes
Nrl-Rho	x	x			x	yes
Crx-Rho	x	x				yes

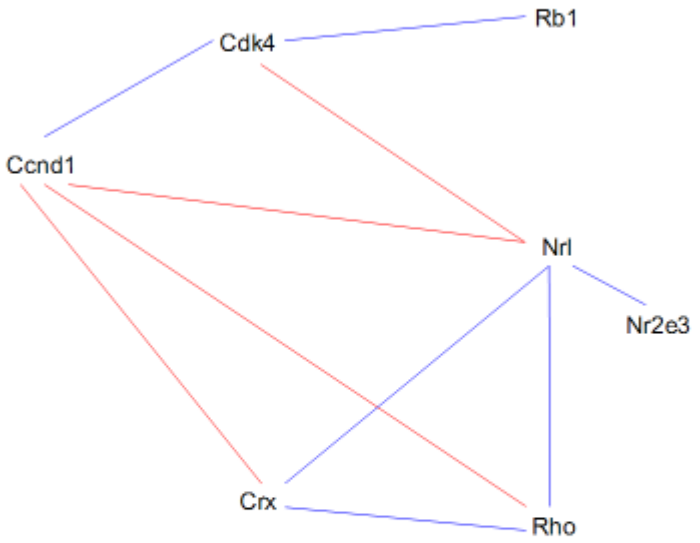

Figure 2 A rod network reconstructed based on correlations among seed genes in the expression datasets. Links were drawn to connect any two seed genes with a correlation of |0.65| or greater in two or more of the five datasets. Blue lines represent positive correlations and red lines represent negative correlations.

degree of overall agreement among the different datasets (with the largest observed correlation of correlations between any pair of datasets being only 0.33), demonstrates the usefulness of combining multiple gene expression datasets for inferring gene networks from gene expression data and increasing the robustness of the resulting conclusions.

Prioritizing experimental targets using seed network and expression data

Based on the success of our attempt to (at least partially) recover the links between genes in the seed network, we proceeded to use the seed network to identify additional genes that are likely to be involved in rod differentiation. To do this we queried the gene expression datasets using a procedure similar to the one we used to reconstruct the seed network. For each of our seed genes, we generated a list of all genes whose expression levels were positively or negatively correlated with the network gene in at least two of the five datasets, with the absolute value of the correlation in each case being at least 0.65. We then sorted each list by the number of datasets in which a candidate gene in the list met the correlation threshold of a 0.65 (with a seed network gene) as well as by the mean value of these correlations across those datasets, thus producing a list of prioritized candidate genes correlated with each seed network gene (data not shown). To further prioritize the candidate genes, we generated a list of genes whose expression levels were positively or negatively correlated with at least two genes of interest (i.e. seed network genes Nrl, Nr2e3, Crx, Rb1, Chx10, Rho and Neurod1), and met the correlation threshold of positive (or negative) 0.65 in at least two datasets. Using this approach we identified 986 genes whose expression levels are positively correlated with more than 2 genes of interest with a correlation coefficient of at least 0.65 (Supp. Table 1). We then retrieved Gene Ontology and KEGG pathway annotations for the genes in this list. Based on this information we found the MAPK signaling, oxidative phosphorylation, purine metabolism, glycolysis, gluconeogenesis, tight junction neuroactive ligand-receptor interaction, calcium signaling, and insulin signaling pathway annotations to be prominently represented in this list (Supp. Table 3a and b). Similarly, we identified 531 genes whose expression levels are negatively correlated with more than 2 genes of interest. Based on retrieval of Gene Ontology

and KEGG pathway annotations for the genes in this list we found the ribosome, MAPK signaling, cell cycle, axon guidance, regulation of actin cytoskeleton, pyrimidine metabolism, focal adhesion and purine metabolism annotations were prominently represented (Supp. Tables 3c and d).

Genes with known links to photoreceptors

Several of the genes whose expression levels were found to be highly positively correlated with multiple genes in the rod seed network (based on analysis of more than one data set) are known to be important for rod photoreceptor function, e.g. phosphodiesterase 6G, cGMP-specific rod gamma, recoverin, rod outer segment membrane protein 1, and phosducin (Supp. Table 2). The fact that our list of candidate genes includes genes that have strong experimental evidence of involvement in rod photoreceptor functions suggests that the other candidate genes that we have identified through our approach of using a seed network to query multiple expression datasets are worthy of careful consideration in the context of rod development.

Expanding the seed network into a hypothesized rod gene network

Based on the lists generated by this analysis we have identified seven genes or groups of genes that are candidates for immediate inclusion into a hypothesized rod gene network, that extends the seed network. These include Uhmk1, Kruppel-like transcription factor-7, Ext1 and other genes involved in heparan sulfate biosynthesis, cystatin C, N-myc downstream regulated genes 3 and 4, Nr1d2, and ROR-alpha (Figure 3). One additional gene, p27Kip, was added to the hypothesized rod gene network based on its interaction with two candidate genes. We also included p27Kip in the hypothesized rod gene network because it inhibits the seed network gene cdk and has been shown to regulate retinal progenitor cell cycle withdrawal (Dyer and Cepko (2001)). U2AF homology motif (UHM) kinase 1, (Uhmk1; also called Kis or Kinase interacting with stathmin), is a serine/threonine kinase that contains an RNA binding motif (Maucuer et al. (1995, 1997)). Uhmk1 is positively correlated with Nrl, Nr2e3, rhodopsin, and Crx and is negatively correlated with NeuroD1. Uhmk1 has been found to bind to and

Figure 3 Expansion of the seed network to include candidate genes. Genes highly correlated with multiple seed network members were considered for inclusion into the original seed network. Based on published experimental evidence, seven candidate genes or gene families (represented by blue ovals) were identified and proposed links were added to the seed network genes (represented by gray ovals). Red arrows indicate a negative relationships between genes, blue arrows a positive relationships. The dashed arrows indicate hypothesized links not yet verified by direct experimental evidence. The box surrounding Nrl, Nr2e3, and rhodopsin indicates seed network genes which are specific to rod photoreceptors. Candidate genes (blue), which have a link to this box are proposed to interact (likely indirectly) with several rod genes.

negatively regulate the cell cycle inhibitor p27Kip (Boehm et al. (2002)), which is involved in regulation of retinal progenitor cell fate. This, together with the observed correlation in Uhmk1s expression with the expression of two well characterized transcription factors that direct photoreceptor cell fate (Crx and Nrl) is highly suggestive of its involvement in rod progenitor cell cycle exit.

Several of the Kruppel-like transcription factors are highly correlated with multiple genes in the rod seed network. The Kruppel-like factors function as repressors or activators of transcription and are good candidates for regulation of genes involved in rod development as they are involved in cell proliferation and differentiation in many tissues including the retina (Otteson et al. (2004)). Kruppel-like transcription factor 7 (Klf7) is highly negatively correlated with Crx and Nrl in multiple datasets. Klf7 is expressed in differentiating cells in the embryonic retina and other parts of the central nervous system (Laub et al. (2001, 2005)). Klf7 knockout mice show downregulation of the cdk inhibitor p27Kip and there is evidence that it directly activates the p27Kip promoter. Klf7 may therefore play a key role in regulating the cell cycle of retinal progenitors.

Several genes involved with heparan sulfate biosynthesis are correlated with the expression of genes in the seed network. Exostoses (multiple) 1 or Ext1 is positively correlated with Nrl, rhodopsin, Nr2e3 and Crx. Ext1 is a glycosyltransferase involved in the synthesis of heparan sulfate and is known to be highly expressed in developing mouse brain (Inatani and Yamaguchi (2003)). Other genes involved in heparan sulfate biosynthesis are also highly correlated with multiple genes in our seed network. These include heparan sulfate (glucosamine) 3-O-sulfotransferase 3B1 which is positively correlated with Nrl, rhodopsin and Nr2e3, beta-1,3-glucuronyltransferase 1 (glucuronosyltransferase P) which is positively correlated with Nrl and rhodopsin, and carbohydrate (chondroitin) synthase 1 which is also positively correlated with Nrl and rhodopsin. A role for heparan sulfate in retinal development has been suggested by studies of its expression and heparan sulfate has been shown to have an effect on several pathways important in development such as the hedgehog and fibroblast growth factor pathways (Cool and Nurcombe (2006); Rubin et al. (2002)).

Cystatin C is positively correlated with Nrl, Nr2e3, Crx, and rhodopsin. Cystatin C is a cysteine protease inhibitor found in many tissues including the retina. Cystatin C RNA and protein expression have been detected in the embryonic and postnatal rodent retina with peak levels of the protein expressed around the time of photoreceptor maturation (Barka and van der Noen (1994); Wassélius et al. (2001)). Recently, Kato et al. (2006) isolated cystatin C from conditioned media of primary neurospheres and demonstrated that addition of cystatin C to embryonic stem cells facilitated the differentiation into cells expressing neural genes. The fact that cystatin C is expressed in the developing retina, is implicated in promoting neuronal cell fate determination, and is correlated with multiple seed network genes makes it a likely candidate for involvement in photoreceptor development.

N-myc downstream regulated gene 3 (Ndrg3) is highly positively correlated with Crx, Nrl, and rhodopsin. Another N-myc downstream regulated gene, Ndrg4 is highly correlated with Nrl in two datasets. Ndrg3 and Ndrg4 are inhibited by N-myc, one of the members of the myc family of protooncogenes. N-myc has been shown to be important in central nervous system development and is thought to play a role in CNS cell proliferation and differentiation (Stanton et al. (1992)). N-myc is highly negatively correlated with Nrl and rhodopsin. N-myc is expressed in the developing retina but not in mature retinal neurons (Hirning et al. 1991). N-myc is inhibited by retinoblastoma (Rb1) and expression of Ndrg3 and Ndrg4 are reduced in the Rb knockout retina (data accessible at NCBI GEO database, accession number GSE1129; http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1129). Therefore Rb1 may be important for inhibition of N-myc during cell fate determination in the retina which in turn increases expression of Ndrg3 and Ndrg4. Ndrg3 and Ndrg4 may promote rod differentiation through enhancement of AP-1 activity as Ndrg4 has been shown to regulate activity of the protein complex (Ohki et al. (2002)). AP-1 binding sites are found in the Nrl promoter region and the promoters of other rod specific genes (Farjo et al. (1993)).

The orphan nuclear receptor Nr1d2 is highly correlated with Crx, Nrl, Nr2e3 and rhodopsin. This gene is a member of the Reverb nuclear receptor subgroup along with Reverb alpha (Nr1d1), which can function as transcriptional silencers and can repress transcriptional activa-

tion by retinoid-related orphan receptor alpha (Nr1f1) and thyroid hormone receptor (Forman et al. (1994)). There is evidence that Nr1d1 interacts with Nr2e3 and Nrl to activate transcription of rhodopsin in the retina (Cheng et al. (2004)). Both Reverb proteins bind to the same core promoter sequence suggesting that Nr1d2 may also be involved in activating transcription of rhodopsin and other rod photoreceptor genes.

Another orphan nuclear receptor highly correlated with the rod seed genes Nrl and Crx was retinoid-related orphan receptor alpha (ROR-alpha). ROR-alpha is a member of the steroid/thyroid hormone receptor superfamily. Interestingly it has recently been shown that Nrl contains a putative ROR-alpha response element and other retinoic acid receptor binding sites in its promoter region and that deletion of these elements decreases retinoic acid induced luciferase activity in Nrl promoter-luciferase constructs (Khanna et al. (2006)). Discovering the ligands for ROR-alpha and the Reverb nuclear receptors could reveal factors important for controlling Nrl expression in developing photoreceptors. Examination of the data extracted from the mouse retina SAGE library (http://itstgp01.med.harvard.edu/retina) suggests that ROR-alpha is more highly expressed in the outer nuclear layer of the retina than retinoic acid receptor alpha (RAR-alpha) and its temporal RNA expression more closely correlates with that of Nrl.

Summary of candidate genes

The information available in literature on the candidate genes summarized above makes them likely candidates for linking with specific genes in the rod seed network (Figure 3). Both Uhmk1 and Klf7 may be involved in rod genesis through regulation of cell cycle progression by negative or positive regulation of p27Kip. The orphan nuclear protein ROR-alpha is linked directly to Nrl based on a putative binding site present in the Nrl promoter region. Nr1d2 is linked to rhodopsin based on its similarities to Nr1d1, a protein that is known to bind to the rhodopsin promoter region. Ndrg 3 and 4, genes involved in heparan sulfate biosynthesis, and cystatin C correlated with several rod genes, and are shown to have links with all rod specific genes.

Recently, efforts to identify members of the photoreceptor transcriptional network used mouse knockouts of Nrl, Nr2e3 and Crx to identify genes that may be regulated by, and therefore primarily downstream of these three key transcription factors (Hsiau et al. (2007)). Of the 628 genes dysregulated genes identified by this study, 174 are present in our list of 1789 genes either positively or negatively correlated with multiple seed network members. Our results are complimentary to this study, as our approach is likely to identify candidates upstream of Crx, Nrl and Nr2e3 as well.

Discussion

Several large-scale gene expression studies of the murine retina have been conducted in an attempt to identify genes important for retinal development (Akimoto et al. (2006); Blackshaw et al. (2004); Dorrell et al. (2004); Liu et al. (2006); Mu et al. (2001); Zhang et al. (2006)). The data from these studies provide useful information about the changes in gene expression during retinal development. However, these studies offer at best only a starting point for functional studies focused on a smaller subset of genes. The relatively low degree of correspondence in terms of pairwise correlations in gene expression across datasets from different studies further complicates the use of multiple datasets to extract a small subset of the genes as good candidates for a role in specific events in retinal development (such as rod photoreceptor genesis).

Against this background, we have explored a novel approach for analysis of multiple gene expression datasets to identify genes that are likely to play important roles in rod photoreceptor development. We have demonstrated a simple approach to leveraging multiple gene expression datasets to increase the robustness of inferred links between genes, by focusing on links supported by multiple gene expression datasets. We then used a similar approach to query multiple gene expression datasets, using a seed network consisting of a small number of genes (known to be important in rod development), to identify genes whose expression levels are highly correlated with those of the seed network genes in multiple datasets.

The simple approach to combining information from multiple gene expression datasets used

here does not assign different weights to the evidence provided by the different datasets. It might be useful to consider more robust approaches to leveraging information from multiple gene expression datasets e.g. using a machine learning algorithm (Baldi and Brunak (2001)) to learn the weights to be used to combine the evidence provided by the different datasets in support of links between seed network genes and other genes in the datasets. For example, the weights could be optimized using machine learning so as to maximize the accuracy of reconstruction of the seed network from the available data. The resulting weights could then be used in expanding the seed network by adding new links based on evidence from multiple datasets.

The hypothesized rod network described here summarizes our first results obtained using the approach developed in this paper for querying multiple gene expression datasets using a seed network. Our analysis has focused on narrowing down the list of 986 genes that are positively correlated with at least 2 seed network genes. We have not yet analyzed the list of 531 genes that are negatively correlated with at least 2 seed network genes. Of particular interest are genes that are positively correlated with some seed network genes and negatively correlated with other seed network genes. We have relied mostly on the analysis of Gene Ontology and KEGG pathway annotations of genes that are correlated with at least 2 seed network genes in the broader context of the current literature on retinal development. Several additional sources of information can be brought to bear on the task of further refining the hypothesized rod gene network, e.g. protein-protein interaction data, phosphorylation data, among others. Work in progress is aimed at exploring some of these directions.

Related Work

Several previous studies have examined ways of extending a known seed network (Bader (2003); Cabusora et al. (2005); Can et al. (2005); Dougherty et al. (2000); Hashimoto et al. (2004); Shmulevich et al. (2002)). Most of these focus on filtering or selecting candidate links based on some criteria (Bader (2003); Cabusora et al. (2005); Dougherty et al. (2000); Hashimoto et al. (2004); Shmulevich et al. (2002)) or producing a single ranking of all genes

in terms of the degree to which they are related to the entire seed network (Can et al. (2005)). In contrast, we focus on producing a ranking for each seed gene as well as a ranking of those genes that are correlated with multiple seed genes. The latter is especially useful in showing, at a glance, the specific genes in the seed network that are likely to be involved in interactions with a candidate gene. The resulting prioritized list can then be further examined by human experts in the broader context of related literature and biological knowledge.

Summary

By using a seed network to query multiple retinal gene expression datasets we were able to identify candidate genes for further study related to rod photoreceptor development. We used the seed network to prioritize genes in the datasets based on their correlation with multiple seed network members. Based on further analysis of the prioritized lists in the context of evidence obtained from the literature in support of the new links, we were able to identify a small subset of genes from the prioritized lists for addition to the seed network. These new links in the resulting rod gene network offer a rich source of hypotheses that can help focus the experiments at the bench. We believe that this approach offers a powerful means of leveraging computational analysis of high-throughput gene expression data, together with the interpretation of the results by biologists in the context of existing biological knowledge, to rapidly identify and prioritize experimental targets.

Acknowledgements

This work was supported in part by a grant from the National Institutes of Health: EY014931 and a USDA-Food and Agricultural Sciences-Multidisciplinary Graduate Education and Training Grant: 2001-52100-11506.

GENERAL CONCLUSIONS

Summary

We have explored here a simple novel approach to combining multiple datasets across different studies and different platforms and then using the combined data to prioritize genes for further inspection and targeted experimentation. Many thousands of gene expression datasets exist, so it would be of enormous benefit to researchers to have a simple effective procedure for comparison and integration of multiple datasets relevant to a particular topic of interest. One of the drawbacks of high-throughput methods such as microarray or SAGE is the presence of noise in the measurements, which can cause two genes to appear highly correlated when they are actually not. We hoped that our method of combining these datasets would allow us to reveal which gene correlations were robust across different studies, and our results support that this is in fact the case. Unfortunately even robust gene correlations will often occur by chance when there are only a few conditions or time points for thousands of genes. This is why we chose to extend outward from a seed network of genes known to play a part in influencing retinal progenitor cells to adopt a rod photoreceptor cell fate. By using this seed network to query the datasets, we were able to successfully prioritize genes of interest for further experimentation based on robust correlations with multiple seed network genes.

Suggestions for Future Research

We allowed each dataset to have an equal say in whether a gene correlation was determined to be robust. It could be beneficial to use a machine learning approach to assign different weights to different datasets based on how reliable they are learned to be. One method would

be to optimize the weights to maximize how accurately the seed network is reconstructed. The resulting weights could then be used to expand the seed network by adding new links based on evidence from each dataset.

Also, our methods did not take any advantage of the fact that in a time series, the expression of a gene at a particular time point will be dependent on the expression of that gene at previous timepoints. It might be useful to consider ways of possibly taking advantage of this temporal dependence. One way might be to, instead of using a Spearman rank correlation above a certain threshold as a vote for the existence of a link between two genes, use occurrence of co-clustering via a method designed for time series, such as STEM (Short Time-series Expression Miner) (Ernst and Bar-Joseph (2006)).

We chose in this work to focus on the non-trivial problem of integrating gene expression datasets from different platforms and how to query that integrated data to expand a known seed network. But in addition to gene expression data, there are many other types of high-throughput data that could be used to try to cast light on links between genes, including transcription factor binding data, yeast two-hybrid data, Gene Ontology data, etc. Since each type of data provides a different kind of biological information and each type has its own weaknesses, it would seem beneficial if they could be combined in such a way that they could complement and reinforce each other. Some recent studies have attempted to make use of such additional information (Gunsalus et al. (2005); Rhodes et al. (2005); Xia et al. (2006); Pujana et al. (2007); English and Butte (2007)). Advances in these kinds of approaches have the potential to allow us to infer relationships among genes that may not be discernable from gene expression data alone.

So far, Nrl is the furthest upstream gene found to influence retinal progenitor cells to adopt a rod fate rather than a cone fate, and studies have been done to determine the downstream effects on gene expression of knocking out Nrl (Akimoto et al. (2006); Corbo et al. (2007)). It would be very interesting to know, however, what cell fate cues might exist upstream from Nrl. When Nrl is knocked out, it seems to be the case that those cells that would normally have become rods instead become cone-like in many of their morphological, molecular and

electophysiological features (Daniele et al. (2005)). If the developmental gene expression of these pseudocones could be compared with that of genuine cones, it could help to elucidate differences between rods and cones that are due to factors other than Nrl's presence or absence. Unfortunately, this is very difficult in practice since, in the murine retina at least, the overwhelming majority (97%) of photoreceptors are rods.

Conclusion

We believe that the approach we have presented here offers a powerful first step towards leveraging computational analysis of high-throughput gene expression data, together with the interpretation of the results by biologists in the context of existing biological knowledge, to rapidly identify and prioritize experimental targets.

APPENDIX A. METHODS DETAILS

Matching gene identifiers from different platforms

We chose to map all genes from the five mRNA expression datasets to Entrez Gene IDs, since Entrez is a comprehensive and well-maintained standard. For the three Affymetrix microarray datasets (Akimoto et al. (2006); Dorrell et al. (2004); Liu et al. (2006)), which were measured using two different Affymetrix microarray chips, WebGestalt (Zhang et al. (2005)) was used to map the Affymetrix probe IDs to Entrez Gene IDs. For the cDNA microarray dataset (Zhang et al. (2006)), WebGestalt was used to map the given Unigene IDs to Entrez Gene IDs.

In the case of the SAGE dataset (Blackshaw et al. (2004)), ID mapping was a little trickier. Although the SAGE dataset provided Unigene IDs, a large number of them had already been retired. The SAGE dataset also provided gene symbols, however very many genes, if not most genes, are referred to by more than one gene symbol in the literature, and some gene symbols have even been used to refer to more than one gene, making it problematic to try to uniquely map Entrez Gene IDs from the provided gene symbols. To work around these difficulties, We came up with five methods for identifying genes based on the information given in the SAGE dataset: 1) Map the non-retired Unigene IDs via WebGestalt (some of these Unigene IDs were already retired); 2) Map the gene symbols in Blackshaw's data via WebGestalt (some of these gene symbols were ambiguous); 3) Map SAGE tags to Entrez Gene IDs via NCBI SAGEmap (Lash et al. (2000)) (for many tags there are two or three reliable Unigene IDs, and for a number of Unigene IDs there are multiple EntrezGene IDs); 4) Look up Unigene IDS from probe accession numbers given in the ISH results table from Blackshaw et al. (2004) (again, for a number of Unigene IDs there are multiple EntrezGene IDs); and 5) Using the gene names

listed in the tables to disambiguate cases where there are multiple possibilities (methods 3 and 4 above).

Once we completed my mapping of Blackshaw's SAGE tags to gene symbols and Entrez-Gene IDs using each of the five methods outlined above, we found that there was fortunately relatively little disagreement between methods. If there were any differences between methods in the mapping obtained for a particular SAGE tag, then that tag was omitted from this study unless there was a clear majority of evidence for a particular gene mapping.

APPENDIX B. SUPPLEMENTARY DATA

Supplemental tables from the journal article

Supplementary Table 1

Genes that correlate with multiple seed genes (correlation value of 0.65 or greater in at least two datasets) are listed. A correlation of 0 in this table indicates that the gene was not present in a particular dataset.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735966/bin/Supp_table_1.xls

Supplementary Table 2

This contains the subset of genes from Supplementary Table 1 that are expressed in photoreceptors. For each gene that is listed, the correlated seed gene is indicated as well as the mean correlation across datasets in which the correlation reached threshold.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735966/bin/Supp_table_2.pdf

Supplementary Table 3a

This table lists the number of times a KEGG annotation was retrieved using the list of genes positively correlated with multiple rod seed network genes.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735966/bin/Supp_table_3a.xls

Supplementary Table 3b

This list contains genes positively correlated with multiple seed network genes that also have an annotation linking them to a pathway. Genes are listed by their Unigene symbol and

are grouped according to the signaling pathways with which they are associated.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735966/bin/Supp_table_3b.xls

Supplementary Table 3c

This table lists the number of times a KEGG annotation was retrieved using the list of genes negatively correlated with multiple rod seed network genes.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735966/bin/Supp_table_3c.xls

Supplementary Table 3d

This list contains genes negatively correlated with multiple seed network genes that also have an annotation linking them to a pathway. Genes are listed by their Unigene symbol and are grouped according to the signaling pathways with which they are associated.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735966/bin/Supp_table_3d.xls

BIBLIOGRAPHY

Adler, R. and Raymond, P. A. (2008). Have we achieved a unified model of photoreceptor cell fate specification in vertebrates? *Brain Res*, 1192:134–150.

Ahmad, I., Acharya, H. R., Rogers, J. A., Shibata, A., Smithgall, T. E., and Dooley, C. M. (1998). The role of neurod as a differentiation factor in the mammalian retina. *J Mol Neurosci*, 11(2):165–178.

Akimoto, M., Cheng, H., Zhu, D., Brzezinski, J. A., Khanna, R., Filippova, E., Oh, E. C. T., Jing, Y., Linares, J.-L., Brooks, M., Zareparsi, S., Mears, A. J., Hero, A., Glaser, T., and Swaroop, A. (2006). Targeting of gfp to newborn rods by nrl promoter and temporal expression profiling of flow-sorted photoreceptors. *Proc Natl Acad Sci U S A*, 103(10):3890–3895.

Asthana, S., King, O. D., Gibbons, F. D., and Roth, F. P. (2004). Predicting protein complex membership using probabilistic network reliability. *Genome Res*, 14(6):1170–1175.

Bader, J. S. (2003). Greedily building protein networks with confidence. *Bioinformatics*, 19(15):1869–1874.

Baldi, P. and Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach*. MIT Press.

Barka, T. and van der Noen, H. (1994). Expression of the cysteine proteinase inhibitor cystatin c mrna in rat eye. *Anat Rec*, 239(3):343–348.

Blackshaw, S., Fraioli, R. E., Furukawa, T., and Cepko, C. L. (2001). Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes. *Cell*, 107(5):579–589.

Blackshaw, S., Harpavat, S., Trimarchi, J., Cai, L., Huang, H., Kuo, W. P., Weber, G., Lee, K., Fraioli, R. E., Cho, S.-H., Yung, R., Asch, E., Ohno-Machado, L., Wong, W. H., and Cepko, C. L. (2004). Genomic analysis of mouse retinal development. *PLoS Biol*, 2(9):E247.

Boehm, M., Yoshimoto, T., Crook, M. F., Nallamshetty, S., True, A., Nabel, G. J., and Nabel, E. G. (2002). A growth factor-dependent nuclear kinase phosphorylates p27(kip1) and regulates cell cycle progression. *EMBO J*, 21(13):3390–3401.

Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett*, 573(1-3):83–92.

Cabusora, L., Sutton, E., Fulmer, A., and Forst, C. V. (2005). Differential network expression during drug and stress response. *Bioinformatics*, 21(12):2898–2905.

Cai, L., Huang, H., Blackshaw, S., Liu, J. S., Cepko, C., and Wong, W. H. (2004). Clustering analysis of sage data using a poisson approach. *Genome Biol*, 5(7):R51.

Can, T., Camoglu, O., and Singh, A. (2005). Analysis of protein-protein interaction networks using random walks. In *Conference on Knowledge Discovery in Data*.

Chacko, D. M., Rogers, J. A., Turner, J. E., and Ahmad, I. (2000). Survival and differentiation of cultured retinal progenitors transplanted in the subretinal space of the rat. *Biochem Biophys Res Commun*, 268(3):842–846.

Chen, C. M. and Cepko, C. L. (2000). Expression of chx10 and chx10-1 in the developing chicken retina. *Mech Dev*, 90(2):293–297.

Chen, S., Wang, Q. L., Nie, Z., Sun, H., Lennon, G., Copeland, N. G., Gilbert, D. J., Jenkins, N. A., and Zack, D. J. (1997). Crx, a novel otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. *Neuron*, 19(5):1017–1030.

Cheng, H., Khanna, H., Oh, E. C. T., Hicks, D., Mitton, K. P., and Swaroop, A. (2004).

Photoreceptor-specific nuclear receptor nr2e3 functions as a transcriptional activator in rod photoreceptors. *Hum Mol Genet*, 13(15):1563–1575.

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19 Suppl 1:i84–i90.

Claverie, J. M. (1999). Computational methods for the identification of differential and coordinated gene expression. *Hum Mol Genet*, 8(10):1821–1832.

Conlon, E. M., Song, J. J., and Liu, A. (2007). Bayesian meta-analysis models for microarray data: a comparative study. *BMC Bioinformatics*, 8:80.

Conlon, E. M., Song, J. J., and Liu, J. S. (2006). Bayesian models for pooling microarray studies with multiple sources of replications. *BMC Bioinformatics*, 7:247.

Cool, S. M. and Nurcombe, V. (2006). Heparan sulfate regulation of progenitor cell fate. *J Cell Biochem*, 99(4):1040–1051.

Corbo, J. C., Myers, C. A., Lawrence, K. A., Jadhav, A. P., and Cepko, C. L. (2007). A typology of photoreceptor gene expression patterns in the mouse. *Proc Natl Acad Sci U S A*, 104(29):12069–12074.

Daniele, L. L., Lillo, C., Lyubarsky, A. L., Nikonov, S. S., Philp, N., Mears, A. J., Swaroop, A., Williams, D. S., and Pugh, E. N. (2005). Cone-like morphological, molecular, and electrophysiological features of the photoreceptors of the nrl knockout mouse. *Invest Ophthalmol Vis Sci*, 46(6):2156–2167.

Diaz, E., Yang, Y. H., Ferreira, T., Loh, K. C., Okazaki, Y., Hayashizaki, Y., Tessier-Lavigne, M., Speed, T. P., and Ngai, J. (2003). Analysis of gene expression in the developing mouse retina. *Proc Natl Acad Sci USA*, 100:5491–5496.

Dorrell, M. I., Aguilar, E., Weber, C., and Friedlander, M. (2004). Global gene expression analysis of the developing postnatal mouse retina. *Invest Ophthalmol Vis Sci*, 45(3):1009–1019.

Dougherty, E. R., Kim, S., and Chen, Y. (2000). Coefficient of determination in nonlinear signal processing. *Signal Processing*, 80:2219–2235.

Dyer, M. A. and Cepko, C. L. (2001). p27kip1 and p57kip2 regulate proliferation in distinct retinal progenitor cell populations. *J Neurosci*, 21(12):4259–4271.

English, S. B. and Butte, A. J. (2007). Evaluation and integration of 49 genome-wide experiments and the prediction of previously unknown obesity-related genes. *Bioinformatics*, 23(21):2910–2917.

Ernst, J. and Bar-Joseph, Z. (2006). Stem: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, 7:191.

Farjo, Q., Jackson, A. U., Xu, J., Gryzenia, M., Skolnick, C., Agarwal, N., and Swaroop, A. (1993). Molecular characterization of the murine neural retina leucine zipper gene, nrl. *Genomics*, 18(2):216–222.

Forman, B. M., Chen, J., Blumberg, B., Kliewer, S. A., Henshaw, R., Ong, E. S., and Evans, R. M. (1994). Cross-talk among ror alpha 1 and the rev-erb family of orphan nuclear receptors. *Mol. Endocrinol.*, 8:1253–1261.

Furukawa, A., Koike, C., Lippincott, P., Cepko, C. L., and Furukawa, T. (2002). The mouse crx 5'-upstream transgene sequence directs cell-specific and developmentally regulated expression in retinal photoreceptor cells. *J Neurosci*, 22(5):1640–1647.

Furukawa, T., Morrow, E. M., and Cepko, C. L. (1997). Crx, a novel otx-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation. *Cell*, 91(4):531–541.

Green, E. S., Stubbs, J. L., and Levine, E. M. (2003). Genetic rescue of cell number in a mouse model of microphthalmia: interactions between chx10 and g1-phase cell cycle regulators. *Development*, 130(3):539–552.

Griffith, O. L., Pleasance, E. D., Fulton, D. L., Oveisi, M., Ester, M., Siddiqui, A. S., and Jones, S. J. M. (2005). Assessment and integration of publicly available sage, cdna microarray, and oligonucleotide microarray expression data for global coexpression analyses. *Genomics*, 86(4):476–488.

Gunsalus, K. C., Ge, H., Schetter, A. J., Goldberg, D. S., Han, J.-D. J., Hao, T., Berriz, G. F., Bertin, N., Huang, J., Chuang, L.-S., Li, N., Mani, R., Hyman, A. A., Sönnichsen, B., Echeverri, C. J., Roth, F. P., Vidal, M., and Piano, F. (2005). Predictive models of molecular machines involved in caenorhabditis elegans early embryogenesis. *Nature*, 436(7052):861–865.

Hashimoto, R. F., Kim, S., Shmulevich, I., Zhang, W., Bittner, M. L., and Dougherty, E. R. (2004). Growing genetic regulatory networks from seed genes. *Bioinformatics*, 20(8):1241–1247.

Haverty, P. M., Hsiao, L.-L., Gullans, S. R., Hansen, U., and Weng, Z. (2004). Limited agreement among three global gene expression methods highlights the requirement for non-global validation. *Bioinformatics*, 20(18):3431–3441.

Hoffelen, S. J. V., Young, M. J., Shatos, M. A., and Sakaguchi, D. S. (2003). Incorporation of murine brain progenitor cells into the developing mammalian retina. *Invest Ophthalmol Vis Sci*, 44(1):426–434.

Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., and Chory, J. (2006). Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–2827.

Hsiau, T. H.-C., Diaconu, C., Myers, C. A., Lee, J., Cepko, C. L., and Corbo, J. C. (2007). The cis-regulatory logic of the mammalian photoreceptor transcriptional network. *PLoS One*, 2(7):e643.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314.

Inatani, M. and Yamaguchi, Y. (2003). Gene expression of ext1 and ext2 during mouse brain development. *Brain Res Dev Brain Res*, 141(1-2):129–136.

Kato, T., Heike, T., Okawa, K., Haruyama, M., Shiraishi, K., Yoshimoto, M., Nagato, M., Shibata, M., Kumada, T., Yamanaka, Y., Hattori, H., and Nakahata, T. (2006). A neurosphere-derived factor, cystatin c, supports differentiation of es cells into neural stem cells. *Proc Natl Acad Sci U S A*, 103(15):6019–6024.

Khanna, H., Akimoto, M., Siffroi-Fernandez, S., Friedman, J. S., Hicks, D., and Swaroop, A. (2006). Retinoic acid regulates the expression of photoreceptor transcription factor nrl. *J Biol Chem*, 281(37):27327–27334.

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet*, 82(4):949–958.

Kuo, W. P., Jenssen, T.-K., Butte, A. J., Ohno-Machado, L., and Kohane, I. S. (2002). Analysis of matched mrna measurements from two different microarray technologies. *Bioinformatics*, 18(3):405–412.

Lash, A. E., Tolstoshev, C. M., Wagner, L., Schuler, G. D., Strausberg, R. L., Riggins, G. J., and Altschul, S. F. (2000). Sagemap: a public gene expression resource. *Genome Res*, 10(7):1051–1060.

Laub, F., Aldabe, R., Friedrich, V., Ohnishi, S., Yoshida, T., and Ramirez, F. (2001). Developmental expression of mouse krppel-like transcription factor klf7 suggests a potential role in neurogenesis. *Dev Biol*, 233(2):305–318.

Laub, F., Lei, L., Sumiyoshi, H., Kajimura, D., Dragomir, C., Smaldone, S., Puche, A. C., Petros, T. J., Mason, C., Parada, L. F., and Ramirez, F. (2005). Transcription factor klf7 is important for neuronal morphogenesis in selected regions of the nervous system. *Mol Cell Biol*, 25(13):5699–5711.

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Res*, 14(6):1085–1094.

Lee, J. K., Bussey, K. J., Gwadry, F. G., Reinhold, W., Riddick, G., Pelletier, S. L., Nishizuka, S., Szakacs, G., Annereau, J.-P., Shankavaram, U., Lababidi, S., Smith, L. H., Gottesman, M. M., and Weinstein, J. N. (2003). Comparing cdna and oligonucleotide array data: concordance of gene expression across platforms for the nci-60 cancer cells. *Genome Biol*, 4(12):R82.

Li, A. and Horvath, S. (2007). Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics*, 23(2):222–231.

Liu, J., Wang, J., Huang, Q., Higdon, J., Magdaleno, S., Curran, T., and Zuo, J. (2006). Gene expression profiles of mouse retinas during the second and third postnatal weeks. *Brain Res*, 1098(1):113–125.

MacLaren, R. E., Pearson, R. A., MacNeil, A., Douglas, R. H., Salt, T. E., Akimoto, M., Swaroop, A., Sowden, J. C., and Ali, R. R. (2006). Retinal repair by transplantation of photoreceptor precursors. *Nature*, 444(7116):203–207.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2007). Entrez gene: gene-centered information at ncbi. *Nucleic Acids Res*, 35(Database issue):D26–D31.

Marot, G., Foulley, J.-L., Mayer, C.-D., and Jaffrézic, F. (2009). Moderated effect size and p-value combinations for microarray meta-analyses. *Bioinformatics*, 25(20):2692–2699.

Maucuer, A., Camonis, J. H., and Sobel, A. (1995). Stathmin interaction with a putative kinase and coiled-coil-forming protein domains. *Proc Natl Acad Sci U S A*, 92(8):3100–3104.

Maucuer, A., Ozon, S., Manceau, V., Gavet, O., Lawler, S., Curmi, P., and Sobel, A. (1997). Kis is a protein kinase with an rna recognition motif. *J Biol Chem*, 272(37):23151–23156.

Mears, A. J., Kondo, M., Swain, P. K., Takada, Y., Bush, R. A., Saunders, T. L., Sieving, P. A., and Swaroop, A. (2001). Nrl is required for rod photoreceptor development. *Nat Genet*, 29(4):447–452.

Morrow, E. M., Belliveau, M. J., and Cepko, C. L. (1998). Two phases of rod photoreceptor differentiation during rat retinal development. *J Neurosci*, 18(10):3738–3748.

Mu, X., Zhao, S., Pershad, R., Hsieh, T. F., Scarpa, A., Wang, S. W., White, R. A., Beremand, P. D., Thomas, T. L., Gan, L., and Klein, W. H. (2001). Gene expression in the developing mouse retina by est sequencing and microarray analysis. *Nucleic Acids Res*, 29(24):4983–4993.

Nishida, A., Furukawa, A., Koike, C., Tano, Y., Aizawa, S., Matsuo, I., and Furukawa, T. (2003). Otx2 homeobox gene controls retinal photoreceptor cell fate and pineal gland development. *Nat Neurosci*, 6(12):1255–1263.

Ohki, T., Hongo, S., Nakada, N., Maeda, A., and Takeda, M. (2002). Inhibition of neurite outgrowth by reduced level of ndrg4 protein in antisense transfected pc12 cells. *Brain Res Dev Brain Res*, 135(1-2):55–63.

Otteson, D. C., Liu, Y., Lai, H., Wang, C., Gray, S., Jain, M. K., and Zack, D. J. (2004). Kruppel-like factor 15, a zincfinger transcriptional regulator, represses the rhodopsin and interphotoreceptor retinoid-binding protein promoters. *Invest. Ophthalmol. Vis. Sci.*, 45:2522–2530.

Parmigiani, G., Garrett-Mayer, E. S., Anbazhagan, R., and Gabrielson, E. (2004). A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin Cancer Res*, 10(9):2922–2927.

Pennesi, M. E., Cho, J.-H., Yang, Z., Wu, S. H., Zhang, J., Wu, S. M., and Tsai, M.-J. (2003). Beta2/neurod1 null mice: a new model for transcription factor-dependent photoreceptor degeneration. *J Neurosci*, 23(2):453–461.

Pujana, M. A., Han, J.-D. J., Starita, L. M., Stevens, K. N., Tewari, M., Ahn, J. S., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B., Assmann, V., Elshamy, W. M., Rual, J.-F., Levine, D., Rozek, L. S., Gelman, R. S., Gunsalus, K. C., Greenberg, R. A., Sobhian, B., Bertin, N., Venkatesan, K., Ayivi-Guedehoussou, N., Sol, X., Hernndez, P., Lzaro, C., Nathanson,

K. L., Weber, B. L., Cusick, M. E., Hill, D. E., Offit, K., Livingston, D. M., Gruber, S. B., Parvin, J. D., and Vidal, M. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet*, 39(11):1338–1349.

Rajaram, S. (2009). A novel meta-analysis method exploiting consistency of high-throughput experiments. *Bioinformatics*, 25(5):636–642.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabsi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555.

Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A. M. (2005). Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, 23(8):951–959.

Rubin, J. B., Choi, Y., and Segal, R. A. (2002). Cerebellar proteoglycans regulate sonic hedgehog responses during development. *Development*, 129(9):2223–2232.

Rutherford, A. D., Dhomen, N., Smith, H. K., and Sowden, J. C. (2004). Delayed expression of the crx gene and photoreceptor development in the chx10-deficient retina. *Invest Ophthalmol Vis Sci*, 45(2):375–384.

Sakaguchi, D. S., Hoffelen, S. J. V., Theusch, E., Parker, E., Orasky, J., Harper, M. M., Benediktsson, A., and Young, M. J. (2004). Transplantation of neural progenitor cells into the developing retina of the brazilian opossum: an in vivo system for studying stem/progenitor cell plasticity. *Dev Neurosci*, 26(5-6):336–345.

Sakaguchi, D. S., Hoffelen, S. J. V., and Young, M. J. (2003). Differentiation and morphological integration of neural progenitor cells transplanted into the developing mammalian eye. *Ann N Y Acad Sci*, 995:127–139.

Serebriiskii, I., Estojak, J., Berman, M., and Golemis, E. A. (2000). Approaches to detecting false positives in yeast two-hybrid systems. *Biotechniques*, 28(2):328–30, 332–6.

Shmulevich, I., Dougherty, E. R., Kim, S., and Zhang, W. (2002). Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2):261–274.

Sicinski, P., Donaher, J. L., Parker, S. B., Li, T., Fazeli, A., Gardner, H., Haslam, S. Z., Bronson, R. T., Elledge, S. J., and Weinberg, R. A. (1995). Cyclin d1 provides a link between development and oncogenesis in the retina and breast. *Cell*, 82(4):621–630.

Stanton, B. R., Perkins, A. S., Tessarollo, L., Sassoon, D. A., and Parada, L. F. (1992). Loss of n-myc function results in embryonic lethality and failure of the epithelial component of the embryo to develop. *Genes Dev*, 6(12A):2235–2247.

Stevens, J. R. and Doerge, R. W. (2005). Meta-analysis combines affymetrix microarray results across laboratories. *Comp Funct Genomics*, 6(3):116–122.

Tan, P. K., Downey, T. J., Spitznagel, E. L., Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M., and Cam, M. C. (2003). Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res*, 31(19):5676–5684.

Warnat, P., Eils, R., and Brors, B. (2005). Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 6:265.

Wassélius, J., Håkansson, K., Johansson, K., Abrahamson, M., and Ehinger, B. (2001). Identification and localization of retinal cystatin c. *Invest Ophthalmol Vis Sci*, 42(8):1901–1906.

Weinberg, R. A. (1995). The retinoblastoma protein and cell cycle control. *Cell*, 81(3):323–330.

Wolfe, C. J., Kohane, I. S., and Butte, A. J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, 6:227.

Xia, K., Dong, D., and Han, J.-D. J. (2006). Intnetdb v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. *BMC Bioinformatics*, 7:508.

Yu, J., Farjo, R., MacNee, S. P., Baehr, W., Stambolian, D. E., and Swaroop, A. (2003). Annotation and analysis of 10,000 expressed sequence tags from developing mouse eye and adult retina. *Genome Biol*, 4(10):R65.

Zhang, B., Kirov, S., and Snoddy, J. (2005). Webgestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res*, 33(Web Server issue):W741–W748.

Zhang, J., Gray, J., Wu, L., Leone, G., Rowan, S., Cepko, C. L., Zhu, X., Craft, C. M., and Dyer, M. A. (2004). Rb regulates proliferation and rod photoreceptor development in the mouse retina. *Nat Genet*, 36(4):351–360.

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

Zhou, X. and Mao, K. Z. (2006). Regularization network-based gene selection for microarray data analysis. *Int J Neural Syst*, 16(5):341–352.

Zhou, X. J., Kao, M.-C. J., Huang, H., Wong, A., Nunez-Iglesias, J., Primig, M., Aparicio, O. M., Finch, C. E., Morgan, T. E., and Wong, W. H. (2005). Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat Biotechnol*, 23(2):238–243.