Genomics Data 2 (2014) 369–374

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Development of novel filtering criteria to analyze RNA-sequencing data obtained from the murine ocular lens during embryogenesis

Abby L. Manthey [a], Anne M. Terrell [a], Salil A. Lachke [a], Shawn W. Polson [b], Melinda K. Duncan [a,*]

[a] Department of Biological Sciences, University of Delaware, Newark, DE, USA
[b] Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA

ARTICLE INFO

Article history:
Received 11 October 2014
Accepted 15 October 2014
Available online 24 October 2014

Keywords:
Lens
RNAseq
Biological relevance
Filtering
Embryo

ABSTRACT

Next-generation sequencing of the transcriptome (RNA-Seq) is a powerful method that allows for the quantitative determination of absolute gene expression, and can be used to investigate how these levels change in response to an experimental manipulation or disease condition. The sensitivity of this method allows one to analyze transcript levels of all expressed genes, including low abundance transcripts that encode important regulatory molecules, providing valuable insights into the global effects of experimental manipulations. However, this increased sensitivity can also make it challenging to ascertain which expression changes are biologically significant. Here, we describe a novel set of filtering criteria – based on biological insights and computational approaches – that were applied to prioritize genes for further study from an extensive number of differentially expressed transcripts in lenses lacking Smad interacting protein 1 (Sip1) obtained via RNA-Seq by Manthey and colleagues in *Mechanisms of Development* (Manthey et al., 2014). Notably, this workflow allowed an original list of over 7100 statistically significant differentially expressed genes (DEGs) to be winnowed down to 190 DEGs that likely play a biologically significant role in Sip1 function during lens development. Focusing on genes whose expression was upregulated or downregulated in a manner opposite to what normally occurs during lens development, we identified 78 genes that appear to be strongly dependent on Sip1 function. From these data (GEO accession number GSE49949), it appears that Sip1 regulates multiple genes in the lens that are generally distinct from those regulated by Sip1 in other cellular contexts, including genes whose expression is prominent in the early head ectoderm, from which the lens differentiates. Further, the analysis criteria outlined here represent a filtering scheme that can be used to prioritize genes in future RNA-Seq investigations performed at this stage of ocular lens development.

Specifications	
Organism/cell line/tissue	*Mus musculus; 15.5 day embryonic lens tissue*
Sex	*N/A*
Sequencer or array type	*Illumina HiSeq 2000*
Data format	*Raw (FASTQ) and analyzed (normalized RPKM; SOFT, MINiML, TXT)*
Experimental factors	*Inbred (C57Bl/6<har>) wild type* vs. *Mixed background ($Sip1^{flox/flox}$ no Cre) wild type; mixed background wild type* vs. *Sip1 conditional knockout ($Sip1^{flox/flox}$ + MLR10Cre)*
Experimental features	*Global identification of differentially expressed genes in mice lacking Sip1 in the lens compared to mixed background wild type controls using experimentally derived threshold values, one of which was determined by comparing the lens transcriptomes of inbred and mixed background wild type mice.*
Consent	*N/A*
Sample source location	*Newark, Delaware, USA*

* Corresponding author at: Melinda K. Duncan, Professor, Department of Biological Sciences, University of Delaware, Newark DE 19716.
E-mail address: duncanm@udel.edu (M.K. Duncan).

Direct link to deposited data

Deposited data can be found at: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49949.

Experimental design, materials, and methods

Mouse models and genotyping

All mice in this study were bred and maintained in the University of Delaware Animal Facility and adhered to the Association for Research in Vision and Ophthalmology (ARVO) Statement for the Use of Animals in Ophthalmic and Vision Research. All animal protocols were approved by the University of Delaware Institutional Animal Care and Use Committee (IACUC) (approval number: 1039).

Smad interacting protein 1 (Sip1), a ZEB transcription factor, is expressed as the lens vesicle separates from the head ectoderm in the embryonic mouse [1], becoming more localized to the equatorial lens epithelial cells and transition zone in the adult [2]. An earlier study,

http://dx.doi.org/10.1016/j.gdata.2014.10.015

utilizing LE-Cre to delete the Sip1 gene at the lens placode stage, demonstrated that this gene is important for the separation of the lens vesicle from the presumptive corneal epithelium [3]. However, very little was known about the role of this protein in the lens following this developmental stage. To this end, mixed background mice harboring the *Sip1* gene with exon 7 flanked by LoxP (also known as flox) sites (*Sip1*$^{flox(ex7)}$ or *ZEB2*$^{tm1.1Yhi}$ in the Mouse Genome Informatics Database) [4] were obtained from Dr. Yujiro Higashi (Osaka University, Osaka, Japan). These mice were then crossed to *MLR10Cre* mice expressing Cre recombinase in all lens cells from the lens vesicle stage onward [5], which were originally obtained from Dr. Michael Robinson (Miami University, Oxford, Ohio) on an FVB/N genetic background, then backcrossed four generations to C57Bl/6<har> (Harlan Sprague Dawley, Indianapolis, Indiana) in our laboratory. Embryos were staged by designating the day that the vaginal plug was observed in the dam as E0.5.

In order to genotype these mice, DNA was isolated from adult tail biopsies using the PureGene Tissues and Mouse Tail kit (Gentra Systems, Minneapolis, Minnesota) following the manufacturer's instructions. The DNA was quantitated with an ND-1000 UV–Vis Spectrophotometer (Nanodrop Technologies; Software V3.1.2) and stored at 4 °C until use. Genotyping PCR reactions were done using the following recipe per sample: 10 μl Taq PCR mix (Qiagen, Valencia, California), 1 μl forward primer, 1 μl reverse primer, 7 μl nuclease free water (IDT, Coralville, Indiana), and 1 μl of isolated DNA (approximately 100 ng). Mice were genotyped for the presence of the floxed *Sip1* alleles as well as the *MLR10Cre* transgene using previously described primer sets [1,5] and the following PCR parameters: 30 cycles of 94 °C for 30 s, 58 °C for 30 s, and 72 °C for 30 s, with a final extension at 72 °C for 10 min. Gel electrophoresis was used to separate bands on a 2% agarose ethidium bromide gel followed by visualization on a Carestream Gel Logic 212 Pro.

Following several rounds of mating, we obtained mice with the *Sip1* conditional knockout (cKO) genotype (*Sip1*$^{flox/flox}$ + *MLR10Cre*), which lacked both alleles of the *Sip1* gene from the lens starting at E10.5 onward [1]. These mice had major defects in lens fiber cell migration during development, ultimately leading to cataract formation in the adult. Unfortunately, a candidate gene approach to determine the transcriptional changes responsible for these gross morphological alterations was not fruitful [1]. Thus, we sought to use an unbiased approach to determine the global changes in the transcriptome of *Sip1* cKO lenses at E15.5, which represents the developmental stage immediately proximal to the onset of the most obvious morphological change in these mice.

Notably, at the time this study was performed, no RNA-Seq data had been previously reported for the ocular lens. Although this highlights the novelty of this study, it also posed a question of applicability of this method to such a biased transcriptome as that found in the lens, where the expression of structural genes, such as crystallins, predominates over genes that regulate cell function and phenotype [6]. Therefore, we first examined the ability of RNA-Seq to quantitate the expression of both structural and regulatory genes in E15.5 lenses obtained from inbred mice (C57Bl/6<har>, denoted "AG" in the GEO datasets). These data were also used to estimate the expression variance observed between biological replicates obtained from this inbred strain, as such animals are expected to lack genetic variability outside of that conferred by the sex chromosomes. We further expanded this investigation to include an RNA-Seq analysis of the gene expression differences in phenotypically normal E15.5 lenses arising solely from genetic background variability. To do so, gene expression was compared between lenses isolated from a genetically uniform inbred strain (C57Bl/6<har>; the AG data set) to that observed for embryos homozygous for the *Sip1*^{flox} allele (but lacking Cre), which have a mixed genetic background derived from 129/Sv, C57Bl/6, and FVB/N strains. These *Sip1*$^{flox/flox}$ no Cre animals (denoted "WT" in the GEO datasets) are the source of the wild type controls used in both the phenotypic and gene expression analyses of the *Sip1* cKO lenses, which have a similar mixed genetic background [1].

Sample collection, RNA Isolation, and RNA quality control

Lenses were collected from E15.5 *Sip1* cKO (30 lenses per biological replicate), mixed background wild type (30 lenses per biological replicate), and inbred wild type mice (75 lenses per biological replicate) using micro-dissection, during which the retina, blood vessels, and cornea were carefully removed with forceps. Three biological replicates were collected for each genotype. Total RNA was extracted and isolated using the SV Total RNA Isolation System (Invitrogen, Grand Island, New York) according to the manufacturer's instructions.

To determine the quality and concentration of the isolated RNA, small aliquots of each sample were run on an Agilent 2100 Bioanalyzer using the Agilent RNA 6000 Nano Kit as per the manufacturer's instructions. The RNA integrity number (RIN) and concentration were determined for each sample in order to determine the overall quality and suitability for RNA-Seq. According to the Illumina® TruSeq™ RNA Sample Preparation Kit v2 used for library preparation, a RIN greater than or equal to 8 along with a concentration of at least 0.1 μg is sufficient for mammalian RNA-Seq experiments [7]. Total RNA samples that were of high enough quality were then used for library construction and cluster generation. In this analysis, all samples had RIN values greater than 9.2 and concentrations over 300 ng/μl.

Library preparation and sequencing

Using poly-T oligo attached magnetic beads, mRNA was purified from the total RNA samples and converted to a library of template molecules for cluster generation and DNA sequencing at Global Biologics (Columbia, Missouri) according to the Illumina® TruSeq™ RNA Sample Preparation Kit v2. Briefly, the purified mRNA was mixed with a solution of divalent cations, and then denatured (65 °C), eluted (80 °C), and fragmented. The RNA fragments were then copied into first strand cDNA using reverse transcriptase and random primers, followed by second strand cDNA synthesis using DNA polymerase I and RNase H. The cDNA overhangs resulting from fragmentation were then converted into blunt ends, and the 3′ ends were adenylated with a single nucleotide base to prevent the fragments from ligating to each other and to provide a hybridization target for the adapters, which have a single thymine residue at the 3′ end. Using PCR, the purified, ligated cDNA products were then enriched to create the final cDNA library.

Each of the adaptor-tagged, single-end cDNA libraries was then bound at both ends to a TruSeq v3 flow cell, forming single strand bridges, which were then amplified by binding single molecules to each strand, forming double-stranded bridges. Each bridge was then denatured to form two copies of covalently bound single-stranded template. This was continued to generate a "cluster" of identical copies. Finally, the reverse strands were cleaved and washed away, leaving only the forward strands. The free 3′ ends of these single strands of DNA were blocked and the sequencing primer was hybridized to each. The resulting cDNA library cluster was then sequenced using the SBS Sequencing Kit on an Illumina HiSeq 2000 Sequencer (University of Delaware Genotyping and Sequencing Center) with 50-cycle single-end (50 bp) reads. Using the Illumina Pipeline software (version RTA 1.13.48/CASAVA 1.8.2), the images were analyzed, and the bases called and translated to generate FASTQ sequence files.

Gene mapping and normalization

Next generation sequencing platforms, such as the Illumina HiSeq, produce tens to hundreds of millions of sequence reads during an RNA-Seq experiment. These large volumes of data can obscure evidence of issues that are introduced during library preparation and sequencing. For this reason, a critical first step in RNA-Seq data analysis is to apply

algorithms to examine the data quality. This study utilized both the FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and components of the CLC Genomics Workbench (ver. 6.1, CLC Bio, Aarhus, Denmark) to examine the data for nucleotide usage skews, low per base quality, assignment of ambiguous nucleotides, and unusual sequence content (e.g., overrepresented kmers and sequencing adapters). Additional quality control metrics were also assessed throughout the analysis process to insure the integrity of the sequencing data and the bioinformatic analysis.

Upon confirmation of sequence quality, sequences were trimmed to remove the Illumina TruSeq adapters and poly-A as well as low quality sequence ends (ambiguous base limit: 0, quality limit: 0.01) using the CLC Genomics Server (v. 5.1) Trim Sequences tool. Following trimming, all sequences shorter than 35 bp were discarded. High quality sequences were aligned to the *Mus musculus* reference genome (Build NCBI-M37.65 Ensembl/MGI annotations) using the CLC RNA-Seq reference mapping algorithm (length coverage: 0.9, identity: 0.8). Reads mapping uniquely to exonic portions of an annotated gene were included in observed count totals on a per gene basis. The number of mapped reads per kilobase of transcript per million mapped reads (reads per kilobase per million; RPKM) was calculated from raw counts to rank the expressed genes. Differentially expressed genes (DEGs) were identified by calculating the variance from a beta binomial distribution using the method of Baggerly et al. [8] against quantile normalized observed counts [reviewed in [9]] producing per gene p-values that were false discovery rate (FDR) corrected for multiple comparison [10]. This analysis resulted in the identification of 7108 genes whose expression was altered significantly in the *Sip1* cKO lens compared to the mixed background WT lenses at a 95% confidence level, corresponding to nearly 30% of the predicted mRNA coding genes in the mouse genome [11]. In order to focus solely on the biologically significant expression changes involved in Sip1 function, we developed a filtering strategy using experimentally derived thresholds that estimate which of these changes are likely to be biologically relevant.

Table 1
Expression levels for representative regulatory and structural genes in the E15.5 inbred mouse lens.

Gene symbol	Gene name	Mean expression (RPKM)	Citation
Regulatory genes			
Bin3	Bridging integrator 3	2.6	[21]
Cdh1	E-cadherin	35.0	[22]
Dnase2B	Deoxyribonuclease II beta	16.9	[23]
Fgfr1	Fibroblast growth factor receptor 1	13.4	[24]
Fgfr2	Fibroblast growth factor receptor 2	9.0	[24]
Fgfr3	Fibroblast growth factor receptor 3	41.4	[24]
Foxe3	Forkhead box E3	136.4	[25]
Itga3	α3-Integrin	9.1	[26]
Itga6	α6-Integrin	27.4	[26]
Itgav	αV-Integrin	8.9	[27]
Itgb1	β1-Integrin	74.6	[28]
Jag1	Jagged 1	66.2	[29]
Loxl1	Lysyl oxidase-like 1	24.3	[30]
MAF	Avian musculoaponeurotic fibrosarcoma AS42 oncogene homolog (c-Maf)	152.7	[31]
NHS	Nance–Horan Syndrome	13.3	[32]
Notch2	Notch gene homolog 2	4.1	[33]
Notch 1	Notch gene homolog 1	2.5	[34]
Pax6	Paired box gene 6	11.2	[35]
Prox1	Prospero-related homeobox 1	97.6	[36]
Six3	Sine oculis-related homeobox 3 homolog	13.8	[37]
Tdrd7	Tudor domain containing 7	362	[38]
Txn1	Thioredoxin 1	66.2	[39]
Zeb1	Zinc finger E-box binding homeobox 1 (δEF1)	3.8	[19]
Zeb2	Zinc finger E-box binding homeobox 2 (Sip1)	5.8	[1]
Structural genes			
Cryba1	βA3/A1 crystallin	23,643.5	[40]
Crygf	γF-crystallin	14,682.3	[41]
Mip	Major intrinsic protein of the eye lens fiber (Aquaporin 0)	3621.6	[42]

Filtering strategy

RNA-Seq provides investigators with the ability to rapidly sequence millions to hundreds of millions of transcripts, allowing for the quantitative determination of relative transcript abundance within an mRNA pool. Notably, this ability to simultaneously detect highly expressed as well as rare transcripts, while useful when examining the range of global gene expression changes, may also result in the identification of DEGs that are present at levels far below that needed to affect the biology of a cell or tissue. Thus, we sought to determine filtering criteria to minimize the consideration of DEGs whose expression levels could be reasonably hypothesized to be below the level necessary to affect cellular function or phenotype, allowing us to focus on those that are most likely to be biologically significant. First, we attempted to extrapolate the likely abundance of each mRNA at the level of a single cell. Although the mRNA content of a cell can vary greatly depending on cell type, as well as other factors [12], it has been estimated that a typical mammalian cell contains approximately 500,000 molecules of mRNA [13]. As RNA-Seq data is often normalized and reported as RPKM, a rough estimate equates 2 RPKM to represent approximately one mRNA molecule per cell. However, because of the cellular heterogeneity of the lens, it is arguable that not all genes considered to be expressed in the lens are actually expressed in each cell. To account for this, we also mined the WT RNA-Seq data for regulatory genes with known roles in lens biology to estimate how much mRNA would be necessary to affect lens biology (see Table 1), and found that the vast majority of genes with known functions in the lens are expressed at levels greater than 2 RPKM. Based on these estimates, we only chose genes with a mean RPKM value greater than two for at least one experimental condition for further analysis. Notably, this threshold also appeared to eliminate signals derived from the minimal amount of contamination from neighboring tissues that is, to some extent, inevitable during murine embryonic lens isolation. For example, the expression of the Krüppel-like transcription factors Klf4 and Klf5, which are abundant in the corneal epithelium [14,15], was detected at 0.48 RPKM and 0.09 RPKM respectively, while those for platelet endothelial cell adhesion molecule 1 (Pecam1), a marker expressed abundantly in blood vessels [16], including the blood vessel network surrounding the embryonic lens (known as tunica vasculosa lentis), were also detected at levels below 2 RPKM in the majority of biological replicates.

These filtering criteria were further validated and refined by considering the effect of genetic background variation on gene expression in the lens by comparing the expression levels for E15.5 lenses isolated from an inbred, and thus genetically uniform, strain (the AG data set) and a more genetically diverse wild type population (the WT data set) that had variable genetic contributions from multiple strains. In total, 1611 known genes (i.e., pseudogenes and unknown/predicted sequences were not included), which were expressed at levels over 2 RPKM in either WT or AG, were found to be significantly different between the two samples at more than a 95% confidence level. Notably, though, the differences in expression level detected between these two datasets were under 2.5 fold approximately 99% of the time (1591 genes out of 1611 genes; see Supplementary Table). Further, none of the 20 genes altered over 2.5 fold have known functions in the lens, but have been predicted to function in a wide array of cellular processes (Table 2). This analysis suggests that changes in mRNA expression above 2.5 fold are most likely biologically/functionally significant, and this threshold can be used to better emphasize the biologically relevant effects of genetic manipulations on the mouse lens.

Application of these filtering criteria (unnormalized RPKM over 2 for either WT or mutant lenses, a change in unnormalized RPKM greater than 2 between WT and mutant lens, fold change greater than 2.5 between WT and mutant lens) to the 7108 genes that were significantly

Table 2
Predicted biological function for genes differentially expressed more than 2.5 fold in E15.5 lenses obtained from mixed background (WT) mice compared to those from an inbred strain (AG).

Gene ID	Gene name	Fold change[a]	FDR *p*-value	WT RPKM means[b]	AG RPKM means[b]	Predicted biological function[c]
Snhg9	Small nucleolar RNA host gene (non-protein coding) 9	**6.56**	2.52E−05	1.14	7.79	Long non-coding RNA
Ccbe1	Collagen and calcium binding EGF domains 1	**5.02**	0	0.56	2.98	ECM remodeling/cell migration
Eif3j	Eukaryotic translation initiation factor 3, subunit J	**4.70**	0	6.09	19.64	Component of the translation initiation complex
Adamts4	A disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 4	**3.69**	5.20E−06	1.02	3.93	Degradation of the cartilage proteoglycan aggrecan
Lars2	Leucyl-tRNA synthetase 2, mitochondrial	**2.90**	9.46E−11	150.92	467.79	Leucine tRNA ligase
Mid1	Midline 1	**2.71**	1.72E−03	6.71	18.49	Protein complex anchoring to microtubules
Tdg	Thymine DNA glycosylase	**2.57**	7.95E−14	4.19	16.11	DNA mismatch repair (G/T)
Rpl5	Ribosomal protein L5	**−2.58**	3.61E−29	228.60	176.68	rRNA maturation; 60S ribosomal subunit formation
Cd59a	CD59a antigen	**−2.65**	6.01E−14	5.57	2.09	Complement-mediated cell lysis regulation
Avp	Arginine vasopressin	**−2.66**	1.56E−08	4.94	1.71	Water retention; blood vessel constriction
Lsm7	LSM7 homolog, U6 small nuclear RNA associated (S. cerevisiae)	**−2.74**	8.54E−10	70.55	62.08	U6 snRNA binding during pre-mRNA splicing
Ccdc117	Coiled-coil domain containing 117	**−2.82**	3.51E−19	26.31	9.68	Unknown
Rpl22l1	Ribosomal protein L22 like 1	**−3.00**	6.19E−14	183.84	71.56	Protein component of the ribosome
Nme2	NME/NM23 nucleoside diphosphate kinase 2	**−3.27**	2.28E−10	174.43	132.82	Non-ATP nucleoside triphosphate synthesis
mt-Atp8	Mitochondrially encoded ATP synthase protein 8	**−3.28**	7.69E−09	2343.43	1146.27	Membrane anchor for ATP synthase
Scg5	Secretogranin V	**−4.37**	3.08E−40	2.71	0.61	Molecular chaperone
Gchfr	GTP cyclohydrolase I feedback regulator	**−4.38**	2.89E−04	2.96	0.68	Mediator of GTP cyclohydrolase inhibition
Zfp580	Zinc finger protein 580	**−5.20**	2.22E−03	3.31	1.07	Endothelial cell proliferation and migration
Xlr3b	X-linked lymphocyte-regulated 3B	**−5.22**	3.62E−12	2.59	0.49	Facilitator of parent-of-origin effects on cognitive function
Hist1h2al	Histone cluster 1, H2al	–?	0.011	10.17	0.71	Core component of nucleosomes

[a] Calculated from the normalized RPKM means.
[b] Unnormalized RPKM means.
[c] Predicted biological function as specified by GeneCards (www.genecards.org).

altered in the E15.5 *Sip1* cKO lens compared to the WT lens revealed 190 unique, DEGs with a high likelihood of being relevant to the function of Sip1 in the lens [1].

Data mining iSyTE to determine the usual expression changes for Sip1 cKO DEGs during lens development

To further investigate the relevance of the 190 *Sip1* cKO DEGs to lens biology and to determine their likelihood of being direct Sip1 targets, we sought to compare these candidate genes with existing lens developmental gene expression data in the web-based resource *iSyTE* (*i*ntegrated *Sy*stems *T*ool for *E*ye gene discovery; http://bioinformatics.udel.edu/Research/iSyTE) [17]. *iSyTE* contains gene expression microarray datasets generated on the Affymetrix Mouse Genome 430 2.0 Array platform for wild type outbred ICR mouse lenses at stages ranging from the lens pit/early lens vesicle (E10.5; onset of Sip1 protein expression in the lens [1]) to the lens vesicle (E11.5) and early lens (E12.5). Moreover, *iSyTE* contains microarray datasets on the same platform for mouse whole embryonic tissue without eyes (whole body, WB) from pooled stages E10.5, E11.5, and E12.5 [17]. All microarray datasets in the *iSyTE* study are deposited in GEO (accession number GSE32334), and their analysis has been described in detail [17]. While the lens datasets provide information on the dynamics of gene expression in the developing lens over this period, a *t*-statistic comparison of the lens dataset at each stage with the WB dataset expands the analysis to include the estimation of whether candidate genes exhibit "lens-enriched" expression compared to the remainder of the body. Here, the *iSyTE* lens microarray database was interrogated to compare how the 190 DEGs identified in the *Sip1* cKO lens normally change in gene expression between the onset of Sip1 expression at E10.5 and at E12.5, when Sip1 protein expression in the lens is robust [1]. Based on this comparison, the DEGs in the *Sip1* cKO lens were binned as follows: (A) significantly higher ($p < 0.05$) in the E10.5 lens when compared to E12.5 lens; (B) significantly lower ($p < 0.05$) in the E10.5 lens compared to E12.5 lens; (C) not significantly expressed (detection *p*-value > 0.05) in either the E10.5 or E12.5 lens; (D) not found in the processed Affymetrix dataset; and (E) expressed at levels that are not significantly different between E10.5 and the E12.5 lens. Application of this specialized filter based on normal gene expression changes in the lens that occurs coincident with the onset of Sip1 expression allowed us to narrow down which DEGs are likely to be under Sip1 control. Indeed, a significant number of genes downregulated in the *Sip1* cKO mutant lens were found to be upregulated in the normal lens as it progresses from E10.5 to E12.5. Conversely, a significant subset of genes that were upregulated in the *Sip1* cKO lens were found to be downregulated in the normal lens as it progresses from E10.5 to E12.5. Thus, based on their dynamic expression pattern in the normal lens, this analysis led to the identification of a subset ($n = 78$; 41%) of the 190 *Sip1* cKO DEGs that we believe are strongly dependent on Sip1 function and warrant further investigation to determine their contribution toward the abnormal lens phenotype of the *Sip1* cKO mice.

Discussion

While a candidate gene approach is a useful method that can be used to elucidate the mechanisms underlying a biological process, this approach is inherently biased, and often fails in scenarios where the complexities of a protein's function are not fully known. Sip1 was first described biochemically as a Smad interacting protein and, as such, its function has been intensely investigated in relationship to Smad/transforming growth factor beta (TGFβ) superfamily regulated processes, such as fibrosis and cancer [18]. However, Sip1 is also expressed during development and plays important roles in the formation of the lens, although it does not appear to regulate the same genes during lens development/wound healing that it does in diseases arising outside of the lens [1,3,19]. Thus, after our candidate gene analysis did not uncover the function of Sip1 in the lens, we performed RNA-Seq on *Sip1* cKO lenses as an unbiased approach to better understand the molecular function of this complex gene during lens development. Notably, the main strength of RNA-Seq (i.e., its ability to quantitate gene expression changes with high sensitivity) can also be considered as one of its major disadvantages, as large numbers of statistically significant changes are detected that must then be prioritized for further study. One approach that is often utilized for such prioritization is to use software that determines which DEGs share common biochemical or biological functions, such as Ingenuity Pathway Analysis (IPA; http://www.ingenuity.com/

products/ipa), Gene Set Association Analysis for RNA-Seq (GSAASeq; http://gsaa.unc.edu), Database for Annotation, Visualization, and Integrated Discovery (DAVID; http://david.abcc.ncifcrf.gov/), Integrated Protein Expression (iProXpress; http://proteininformationresource.org/iproxpress/), and many others. However, the databases underlying these software tools introduce their own bias to the data analysis as the literature and other information sources used by these tools focus heavily on human disease pathways, which may not be relevant in other biological contexts. Further, such programs typically do not take the transcript abundance into consideration, which can result in unjustified attention to low abundance, non-biologically relevant DEGs.

Here, we report an integrated method using computational and biological factors to prioritize genes from an expansive list of statistically significant DEGs identified by RNA-Seq experiments. This approach is particularly useful when available pathway analysis tools fail to identify biologically important changes from a list of statistically significant DEGs. We first developed a filtering strategy to restrict the number of genes under consideration to those likely to be expressed at high enough levels to affect cell biology. Generally, the study of architecturally complex tissues would be expected to make the choice of such a cutoff difficult as each cell type has its own unique transcriptome; however, we believe this is less of a concern in the developing lens as this tissue only has two distinct cell types (epithelial and fiber cells) and one transitional stage (epithelial to fiber cell differentiation). Overall, while it is possible that the 2 RPKM cutoff removes some biologically important genes from further consideration, it is likely that the number of these genes is very low and that the majority of biologically relevant genes are retained in the final list of DEGs.

Unfortunately, after applying this minimal expression level filter to the list of genes differentially expressed between the WT and *Sip1* cKO lenses, the number of candidate genes identified was still too large to provide much biological insight. To address this problem, we first compared the lens transcriptomes of two phenotypically normal mouse populations: 1) an inbred strain, which represents a practical source of genetically identical mice, and 2) a randomly bred mixed strain with similar genetic background composition as the *Sip1* cKO mice, only lacking the Cre recombinase transgene. This analysis revealed that these two groups of phenotypically normal lenses expressed over 7700 genes at levels over 2 RPKM, of which 1611 of the known genes were expressed at significantly different levels between these two strains. Notably though, the vast majority (99%) of these significant changes were less than 2.5 fold, leading us to conclude that this fold change represents an ideal cutoff for removing noise resulting from strain to strain variation. Thus, we only considered changes in gene expression that were greater than this fold change in our WT versus *Sip1* cKO comparative analysis. The 20 genes altered more than 2.5 fold in the AG versus WT analysis were also removed from the *Sip1* cKO DEG list as they were likely due to the effect of the mixed genetic background studied. The application of this second filter yielded a manageable list of 190 biologically relevant genes exhibiting differential expression between the WT and *Sip1* cKO lens. However, this list did not include any of the known Sip1 target genes, and established pathway analysis tools did not yield any obvious insight into Sip1 function in the lens. Therefore, in order to better understand the function of the 190 Sip1-regulated genes during lens development, we investigated their normal expression patterns during the very early stages of lens development using the bioinformatics tool, *iSyTE*.

The *iSyTE* database contains microarray-based datasets that compare the relative expression of genes during the earliest stages of mouse lens development (E10.5–12.5, lens pit through early lens) as well as between the lens and a whole embryonic reference dataset [17]. While the first analysis identifies genes expressed in individual stages of the lens, the comparative analysis of lens datasets with the WB allows identification of genes based on their *enriched* expression in the lens. Based on lens-enriched expression, *iSyTE* has been successfully used to help investigators studying the genetic basis of cataracts prioritize, within a mapped interval, the promising candidate genes that are the most likely causing the observed lens phenotype [17]. In this study, the utility of the *iSyTE* dataset was expanded to compare the normal expression pattern of genes differentially expressed in the *Sip1* cKO lens over the period coincident with the onset of Sip1 expression. This approach was particularly useful to determine which expression changes are likely being directly affected by Sip1, as Sip1 is a transcription factor [2,18] and would thus be expected to directly regulate the transcript levels of its target genes. Integrating the *iSyTE* filter enabled us to focus on 41% of the 190 DEGs in the *Sip1* cKO lens. This led to the identification of not only the candidate genes whose expression needed to be downregulated as the lens progressed in normal development (but were not in the *Sip1* cKO mutants), but also the candidates whose expression needed to be upregulated in normal development (but were not in the *Sip1* cKO mutants). As an analogous example, misexpression of Foxe3, a highly lens epithelium-enriched transcription factor that needs to be downregulated in the fiber cells during differentiation, beyond its normal site of downregulation, results in abnormal continued expression of other epithelial markers [20]. At the same time, this also causes a defect in the upregulation of fiber cell genes that are expressed at this stage of cellular differentiation. These data reinforce the argument that deficiencies in transcription factors that function as repressors and activators are expected to cause alterations in the up- or downregulation of genes in normal development, in turn contributing to the pathogenesis of the tissue in mutant phenotypes. Therefore, DEG datasets that result from deficiencies of such regulatory proteins need to be analyzed with reference to the appropriate developmental stage and tissue context to prioritize important candidates.

In conclusion, the ocular environment changes dramatically as the lens vesicle closes, and understanding the global changes in gene expression as well as the function of individual genes during this stage is of critical importance for understanding lens development. This RNA-Seq analysis performed on lenses lacking Sip1 allowed us to show, for the first time, that an important function of Sip1 is to repress the expression of genes which are found in lens precursor cells, but should turn off during normal lens development [1]. Notably, we also found that this function of Sip1 was recapitulated during the lens wound healing response following cataract surgery in adult lenses, further emphasizing that this role is likely to be very important in the lens [19]. Overall, our various analysis filters allowed us to narrow down an unmanageable list of differentially regulated genes from 7108 candidates to 78 (a reduction of over 91 fold) that are highly relevant to our understanding of Sip1 function in the lens.

Conflict of interest

The authors have no conflicts of interest to declare.

Acknowledgements

We thank Dr. Hisato Kondoh/Dr. Yujiro Higashi and Dr. Michael Robinson for providing the Sip1$^{flox(ex7)}$ and MLR10Cre mice, respectively. Library construction and sequencing were performed with the help of Brewster Kingham at the University of Delaware Sequencing Center and Sean Blake at Global Biologics (Columbia, Missouri). Bioinformatic analysis was supported by the staff and computational infrastructure of the University of Delaware Center for Bioinformatics and Computational Biology Core Facility and the Delaware Biotechnology Institute. This work was supported by the NEI Grant, EY12221 supporting M.K.D.; the University of Delaware Chemistry–Biology Interface (CBI) (which is funded by NIH grant T32GM008550) Program supporting A.L.M.; NEI Grant R01EY021505-01 supporting S.A.L.; and Delaware INBRE (NIH National Institute of General Medical Sciences 2P20GM103446-14) and Delaware EPSCoR (National Science Foundation EPS-081425) grants supporting S.W.P.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2014.10.015.

References

[1] A.L. Manthey, S.A. Lachke, P.G. FitzGerald, R.W. Mason, D.A. Scheiblin, J.H. McDonald, M.K. Duncan, Loss of Sip1 leads to migration defects and retention of ectodermal markers during lens development. Mech. Dev. 131 (2014) 86–110.

[2] A.L. Grabitz, M.K. Duncan, Focus on molecules: Smad interacting protein 1 (Sip1, ZEB2, ZFHX1B). Exp. Eye Res. 101 (2012) 105–106.

[3] A. Yoshimoto, Y. Saigou, Y. Higashi, H. Kondoh, Regulation of ocular lens development by Smad-interacting protein 1 involving Foxe3 activation. Development 132 (2005) 4437–4448.

[4] Y. Higashi, M. Maruhashi, L. Nelles, T. Van de Putte, K. Verschueren, T. Miyoshi, A. Yoshimoto, H. Kondoh, D. Huylebroeck, Generation of the floxed allele of the SIP1 (Smad-interacting protein 1) gene for Cre-mediated conditional knockout in the mouse. Genesis 32 (2002) 82–84.

[5] H. Zhao, Y. Yang, C.M. Rizo, P.A. Overbeek, M.L. Robinson, Insertion of a Pax6 consensus binding site into the alphaA-crystallin promoter acts as a lens epithelial cell enhancer in transgenic mice. Invest. Ophthalmol. Vis. Sci. 45 (2004) 1930–1939.

[6] G. Wistow, The NEIBank project for ocular genomics: data-mining gene expression in human and rodent eye tissues. Prog. Retin. Eye Res. 25 (2006) 43–77.

[7] Illumina, TruSeq® RNA Sample Preparation v2 Guide 15026495 F, San Diego, California. 2014.

[8] K.A. Baggerly, L. Deng, J.S. Morris, C.M. Aldaz, Differential expression in SAGE: accounting for normal between-library variation. Bioinformatics 19 (2003) 1477–1483.

[9] B.M. Bolstad, R.A. Irizarry, M. Astrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19 (2003) 185–193.

[10] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 57 (1995) 289–300.

[11] J.L. Guenet, The mouse genome. Genome Res. 15 (2005) 1729–1740.

[12] S. Islam, U. Kjallquist, A. Moliner, P. Zajac, J.B. Fan, P. Lonnerberg, S. Linnarsson, Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 21 (2011) 1160–1167.

[13] S. Bryant, D.L. Manning, Isolation of messenger RNA. Methods Mol. Biol. 86 (1998) 61–64.

[14] S.K. Swamynathan, J.P. Katz, K.H. Kaestner, R. Ashery-Padan, M.A. Crawford, J. Piatigorsky, Conditional deletion of the mouse Klf4 gene results in corneal epithelial fragility, stromal edema, and loss of conjunctival goblet cells. Mol. Cell. Biol. 27 (2007) 182–194.

[15] D. Kenchegowda, S. Swamynathan, D. Gupta, H. Wan, J. Whitsett, S.K. Swamynathan, Conditional disruption of mouse Klf5 results in defective eyelids with malformed meibomian glands, abnormal cornea and loss of conjunctival goblet cells. Dev. Biol. 356 (2011) 5–18.

[16] G. Ferrari, A.R. Hajrasouliha, Z. Sadrai, H. Ueno, S.K. Chauhan, R. Dana, Nerves and neovessels inhibit each other in the cornea. Invest. Ophthalmol. Vis. Sci. 54 (2013) 813–820.

[17] S.A. Lachke, J.W. Ho, G.V. Kryukov, D.J. O'Connell, A. Aboukhalil, M.L. Bulyk, P.J. Park, R.L. Maas, iSyTE: integrated Systems Tool for Eye gene discovery. Invest. Ophthalmol. Vis. Sci. 53 (2012) 1617–1627.

[18] C. Vandewalle, F. Van Roy, G. Berx, The role of the ZEB family of transcription factors in development and disease. Cell. Mol. Life Sci. 66 (2009) 773–787.

[19] A.L. Manthey, A.M. Terrell, Y. Wang, J.R. Taube, A.R. Yallowitz, M.K. Duncan, The Zeb proteins deltaEF1 and Sip1 may have distinct functions in lens cells following cataract surgery. Invest. Ophthalmol. Vis. Sci. 55 (2014) 5445–5455.

[20] H. Landgren, A. Blixt, P. Carlsson, Persistent FoxE3 expression blocks cytoskeletal remodeling and organelle degradation during lens fiber differentiation. Invest. Ophthalmol. Vis. Sci. 49 (2008) 4269–4277.

[21] A. Ramalingam, J.B. Duhadaway, E. Sutanto-Ward, Y. Wang, J. Dinchuk, M. Huang, P.S. Donover, J. Boulden, L.M. McNally, A.P. Soler, A.J. Muller, M.K. Duncan, G.C. Prendergast, Bin3 deletion causes cataracts and increased susceptibility to lymphoma during aging. Cancer Res. 68 (2008) 1683–1690.

[22] G.F. Pontoriero, A.N. Smith, L.A. Miller, G.L. Radice, J.A. West-Mays, R.A. Lang, Co-operative roles for E-cadherin and N-cadherin during lens vesicle separation and lens epithelial cell survival. Dev. Biol. 326 (2009) 403–417.

[23] S. Nishimoto, K. Kawane, R. Watanabe-Fukunaga, H. Fukuyama, Y. Ohsawa, Y. Uchiyama, N. Hashida, N. Ohguro, Y. Tano, T. Morimoto, Y. Fukuda, S. Nagata, Nuclear cataract caused by a lack of DNA degradation in the mouse eye lens. Nature 424 (2003) 1071–1074.

[24] H. Zhao, T. Yang, B.P. Madakashira, C.A. Thiels, C.A. Bechtle, C.M. Garcia, H. Zhang, K. Yu, D.M. Ornitz, D.C. Beebe, M.L. Robinson, Fibroblast growth factor receptor signaling is essential for lens fiber cell differentiation. Dev. Biol. 318 (2008) 276–288.

[25] O. Medina-Martinez, I. Brownell, F. Amaya-Manzanares, Q. Hu, R.R. Behringer, M. Jamrich, Severe defects in proliferation and differentiation of lens cells in Foxe3 null mice. Mol. Cell. Biol. 25 (2005) 8854–8863.

[26] A. De Arcangelis, M. Mark, J. Kreidberg, L. Sorokin, E. Georges-Labouesse, Synergistic activities of alpha3 and alpha6 integrins are required during apical ectodermal ridge formation and organogenesis in the mouse. Development 126 (1999) 3957–3968.

[27] F.A. Mamuya, Y. Wang, V.H. Roop, D.A. Scheiblin, J.C. Zajac, M.K. Duncan, The roles of alphaV integrins in lens EMT and posterior capsular opacification. J. Cell. Mol. Med. 18 (2014) 656–670.

[28] V.N. Simirskii, Y. Wang, M.K. Duncan, Conditional deletion of beta1-integrin from the developing lens leads to loss of the lens epithelial phenotype. Dev. Biol. 306 (2007) 658–668.

[29] T.T. Le, K.W. Conley, N.L. Brown, Jagged 1 is necessary for normal mouse lens formation. Dev. Biol. 328 (2009) 118–126.

[30] J.L. Wiggs, B. Pawlyk, E. Connolly, M. Adamian, J.W. Miller, L.R. Pasquale, R.I. Haddadin, C.L. Grosskreutz, D.J. Rhee, T. Li, Disruption of the blood-aqueous barrier and lens abnormalities in mice lacking lysyl oxidase-like 1 (LOXL1). Invest. Ophthalmol. Vis. Sci. 55 (2014) 856–864.

[31] B.Z. Ring, S.P. Cordes, P.A. Overbeek, G.S. Barsh, Regulation of mouse lens fiber cell development and differentiation by the Maf gene. Development 127 (2000) 307–317.

[32] K.M. Huang, J. Wu, M.K. Duncan, C. Moy, A. Dutra, J. Favor, T. Da, D. Stambolian, Xcat, a novel mouse model for Nance–Horan syndrome inhibits expression of the cytoplasmic-targeted Nhs1 isoform. Hum. Mol. Genet. 15 (2006) 319–327.

[33] S.S. Saravanamuthu, T.T. Le, C.Y. Gao, R.I. Cojocaru, P. Pandiyan, C. Liu, J. Zhang, P.S. Zelenka, N.L. Brown, Conditional ablation of the Notch2 receptor in the ocular lens. Dev. Biol. 362 (2012) 219–229.

[34] S. Rowan, K.W. Conley, T.T. Le, A.L. Donner, R.L. Maas, N.L. Brown, Notch signaling regulates growth and differentiation in the mammalian lens. Dev. Biol. 321 (2008) 111–122.

[35] O. Shaham, Y. Menuchin, C. Farhy, R. Ashery-Padan, Pax6: a multi-level regulator of ocular development. Prog. Retin. Eye Res. 31 (2012) 351–376.

[36] J.T. Wigle, K. Chowdhury, P. Gruss, G. Oliver, Prox1 function is crucial for mouse lens-fibre elongation. Nat. Genet. 21 (1999) 318–322.

[37] W. Liu, O.V. Lagutin, M. Mende, A. Streit, G. Oliver, Six3 activation of Pax6 expression is essential for mammalian lens induction and specification. EMBO J. 25 (2006) 5383–5395.

[38] S.A. Lachke, F.S. Alkuraya, S.C. Kneeland, T. Ohn, A. Aboukhalil, G.R. Howell, I. Saadi, R. Cavallesco, Y. Yue, A.C. Tsai, K.S. Nair, M.I. Cosma, R.S. Smith, E. Hodges, S.M. Alfadhli, A. Al-Hajeri, H.E. Shamseldin, A. Behbehani, G.J. Hannon, M.L. Bulyk, A.V. Drack, P.J. Anderson, S.W. John, R.L. Maas, Mutations in the RNA granule component TDRD7 cause cataract and glaucoma. Science 331 (2011) 1571–1576.

[39] M.F. Lou, Redox regulation in the lens. Prog. Retin. Eye Res. 22 (2003) 657–682.

[40] W. Ferrini, D.F. Schorderet, P. Othenin-Girard, S. Uffer, E. Heon, F.L. Munier, CRYBA3/A1 gene mutation associated with suture-sparing autosomal dominant congenital nuclear cataract: a novel phenotype. Invest. Ophthalmol. Vis. Sci. 45 (2004) 1436–1441.

[41] J. Graw, N. Klopp, A. Neuhauser-Klaus, J. Favor, J. Loster, Crygf(Rop): the first mutation in the Crygf gene causing a unique radial lens opacity. Invest. Ophthalmol. Vis. Sci. 43 (2002) 2998–3002.

[42] A.B. Chepelinsky, Structural function of MIP/aquaporin 0 in the eye lens; genetic defects lead to congenital inherited cataracts. Handb. Exp. Pharmacol. (2009) 265–297.