James Madison University
JMU Scholarly Commons

Senior Honors Projects, 2010-current | Honors College

Spring 2019

Combining chicken retina RNA-seq data across studies to strengthen biomarker detection

Sarah Szvetecz

Follow this and additional works at: https://commons.lib.jmu.edu/honors201019

 Part of the Biostatistics Commons

Recommended Citation
Szvetecz, Sarah, "Combining chicken retina RNA-seq data across studies to strengthen biomarker detection" (2019). *Senior Honors Projects, 2010-current*. 720.
https://commons.lib.jmu.edu/honors201019/720

This Thesis is brought to you for free and open access by the Honors College at JMU Scholarly Commons. It has been accepted for inclusion in Senior Honors Projects, 2010-current by an authorized administrator of JMU Scholarly Commons. For more information, please contact dc_admin@jmu.edu.

Combining Chicken Retina RNA-Seq Data Across Studies to Strengthen Biomarker Detection

An Honors College Project Presented to

the Faculty of the Undergraduate

College of Science and Mathematics

James Madison University

by Sarah Ann Szvetecz

May 2019

Accepted by the faculty of the Mathematics and Statistics Department, James Madison University, in partial fulfillment of the requirements for the Honors College.

FACULTY COMMITTEE:

Project Advisor: Nusrat Jahan, Ph.D.
Associate Professor, Mathematics and Statistics

Reader: Ling Xu, Ph.D.
Associate Professor, Mathematics and Statistics

Reader: Ray Enke, Ph.D.
Assistant Professor, Biology

Reader: ,
,

HONORS COLLEGE APPROVAL:

Bradley R. Newcomer, Ph.D.,
Dean, Honors College

PUBLIC PRESENTATION

This work is accepted for presentation, in part or in full, at Math and Stat Colloquium on April 22, 2019.

Acknowledgements

I would first like to thank my advisor, Dr. Jahan, for her support and guidance throughout this process. I would like to thank her for believing in me, providing me with advice, and committing a significant amount of her time to help me complete my thesis. In addition, I would like to thank my committee members, Dr. Enke and Dr. Xu, for their patience and support in providing useful advice on my writing. I would especially like to thank Dr. Enke for his assistance in helping me better understand the biology and data used in my project, in addition to his help with the bioinformatics steps. It was extremely rewarding for me to be able to work with faculty in both the statistics and biology department and as a result it helped shape me into a stronger researcher and prepare me for graduate school.

I would also like to thank my family and friends for their support during this process. Their constant encouragement during my time working on this project helped me stay focused and motivated to reach my highest potential.

Contents

List of Figures

List of Tables

Abstract

Various studies have identified the chicken embryo (*Gallus gallus*) as a useful model to study the retinogenesis process in humans. This project uses data from two specific RNA sequencing (RNA-seq) studies to investigate retina developmental biology. These studies are done in two different labs using different protocols, as such they cannot be compared directly. Study 1 contains chicken retina samples from embryonic day 3, 5 and 8; while study 2 has retina samples from embryonic day 8, 16, and 18 of developmental age. We apply a normalization method on both studies to account for differences in the two studies. In this work, we perform gene expression analysis on the transformed data to identify genes that could affect the retina development process.

1 Introduction

The retina is a layer of neuronal tissue located in the back of the eye that senses light and objects and sends these images to the brain, which plays a vital role in our vision. Development of the retina stems from optic vesicles that come from both sides of the neural tube. The optic vesicles form the inside of the optic cup becomes the retina and the outside vesicles become the retinal pigment epithelium. The developmental process of the retina continues with cell divisions and migration. This complicated structure consists of thousands of different cells, falling under six major retinal cell types. These six cells types are: bipolar cells, ganglion cells, horizontal cells, retina amacrine cells, and rod and cone photoreceptors [1].

Research of the retina developmental process is necessary to help investigate genes that could play a critical role in the different major cell groups [2]. Being able to find these genes will ultimately help researchers better understand the developmental process. Understanding the retina developmental process is important because problems in the development process can lead to serious issues that affect vision and can eventually lead to blindness if gone untreated. Some examples of these disorders are: retinal detachment, retinoblastoma, and macular degeneration. It is essential to study the developmental process of retina to help understand and be able to better detect retinal disorders [3]. One approach is to examine the changes in gene expressions throughout the developmental process. RNA sequencing analysis provides a framework for studying this process.

1.1 RNA Sequencing

Ribonucleic acid (RNA) sequencing (RNA-seq) is a widely used technique in genomic research to study gene expressions. This technique identifies the amount and sequences of RNA in each sample using next generation sequencing. This process is done by counting the number of randomly sequenced fragments that are aligned with each gene [5]. The procedure consists of extracting total RNA from sample tissues, then converting target RNA molecules to complimentary DNA (cDNA). cDNAs are fragmented, ligated to sequencing adapters, and sequenced on a high-throughput sequencing platform. Currently, the most commonly used high-throughput sequencing platform is Illumina, which generates tens of millions of sequencing reads per sample. The sequencing reads are then mapped back to a reference genome using splice aware sequence alignment software packages such as TopHat [6] or STAR [7]. Gene expression levels are then estimated between samples using a count-based method of aligned reads to each gene in the genome [8]. Popular software for differential expression testing include CuffDiff, edgeR and DEseq [5]. These statistical models use sample replicates to generate a list of differentially expressed genes and also the false discovery rate (FDR) associated with differential expression.

RNA Sequencing is becoming increasingly popular over microarray technology in terms of genomic research applications. This is because RNA-seq is a more in depth, unbiased way of extracting genomic information. RNA sequence based analysis is very sensitive to the transcript length (long versus short). Long transcripts have more detection power for differentially expressed genes as they carry more mapped reads compared to short transcripts. Therefore, it is a complicated process to compare RNA-seq data from different experiments even if they are studying the same organism under similar experimental conditions.

1.2 Connection between chicken retina to human retina

The chicken embryo (*Gallus gallus*) has been identified as a model organism to study the retinogenesis and developmental biology processes in humans [2], [10]. To study retina developmental process, chicken retina is a more feasible organism as compared to the mouse, because their eye function and use are more alike humans. In addition to their eyes compared to the mouse are much larger in size, which makes developmental studies of the retina easier to conduct [2]. Only a small number of RNA sequence studies have been performed on chicken embryos so far. In this project, data from two specific studies were used to identify a unified impact factor for target genes after combing and comparing their results.

1.3 Data

Goal of our research is to investigate the change in gene expressions over the entire retina developmental process. Two data sets were chosen for this study from the NCBI Gene Expression Omnibus (GEO) database [4]. Both data sets contain RNA-seq data sampled at various days of retina development from chicken embryos. Study 1 collected data at embryonic days E3, E5, and E8. On the other hand, study collected data for embryonic days E8, E16, and E18. A more detailed description of the two studies is below.

Study 1 (GSE 89541)

Delayed neurogenesis with respect to eye growth shapes the pigeon retina for high visual activity [10]

Dr. Rodrigues and colleagues from University of Virginia performed this experiment. This study conducts a comparative transcriptome analysis of pigeon and chicken retinas at embryonic stages E3, E5, and E8. For both birds, their samples were triplicated at each embryonic development stage, totaling 18 samples for the study. This study focused mainly on the comparison of the two species and analyzing the differences in the development in the pigeon's retina. The retinal mRNA transcriptome was sequenced with the Illumina HiSeq2500 platform using TruSeq Rapid SBS Kit and HiSeq Rapid Run mode, with 150 based pair single end (SE) reads. The WASHUC2 assembly was used as the reference genome for the alignment. This generated

a total of 14-24 million reads per sample. For the proposed study, only the chicken's data will be relevant (9 samples).

Study 2 (GSE 65938)

RNA sequencing analysis of the developing chicken retina [2]

This experiment was conducted at James Madison University by Dr. Enke's lab. This study conducts a RNA-seq analysis of the developing chicken retina at embryonic stages E8, E16, and E18 to attempt to characterize the mRNA transcriptome to find genes that are critical for retinal development. The samples were duplicated at each embryonic development stage, with an additional 2 cornea samples on E18, totaling 8 samples. The retinal mRNA transcriptome was sequenced using the Illumina NextSeq500 sequencing platform, with 125 based paired end (PE) reads. The galGal4 genome was used as the reference genome to align these reads. This generated a total of 28.6-72.2 million reads per sample. For the proposed study, only the retina data is used (6 samples).

1.4 Conversion from FASTQ file to Raw Count File in Both Studies

To complete this analysis, we needed both data set files to be in the form of raw gene counts. To gain access to this data we needed to convert all the FASTQ files into raw gene counts before beginning an analysis. This process was completed using bioinformatics applications hosted within the CyVerse cyberinfrastructure platform [15]. First, each run for each sample was uploaded into the CyVerse Discovery Environment (DE) using the sample runs corresponding SRA number. For samples with multiple FASTQ files, the Concatenate Multiple Files [15] application was used to combine them into one FASTQ file each. Next, a FastQC [16] report was run to confirm the quality of FASTQ file reads for each sample. Here you want to make sure the average quality score across all the reads is in check. Include example? HISAT2 was then used to align to reads to the gal5 reference chicken genome. Lastly, featureCounts [17] was used to generate the raw read files from the E3- E8 retina data (study1), as well as the E8-E18 data (study2). This process confirms that each of the raw gene count files were generated using identical steps.

1.5 Statistical Analysis

Several different statistical methods were used in this study. The first step was to use a normalization method to account for the library size depth differences between the two studies. After the data was normalized, a differential analysis was performed between different time points of development to identify genes with significant growth over time. Embryonic days E3 and E18 were considered to be of utmost biological

importance in terms of developmental biology. A differential analysis was performed for the genes between day E3 and day E18 to establish a growth pattern of the genes. Various different multivariate analysis methods, such as cluster analysis and principal component analysis were used on selected genes to detect correlated genes and patterns. Finally, developmental processes were tracked for a few selected genes over 6 embryonic days using the combined data from both studies.

2 Methods

2.1 Normalization Methods

Both data sets were individually normalized using the trimmed mean of M-values method (TMM). This is a scaling normalization method that accounts for the difference in library depth sizes. The package edgeR is used [13]. The edgeR function computes a normalizing factor to calculate appropriate scaling factors that are used to rescale the gene counts by dividing the raw gene counts by each sample specific scaling factor. This process is repeated, and the trimmed mean is calculated by taking the sum of the rescaled gene counts for each run and using the following formula below,

$$log_2(d_j^{TMM}) = \frac{\sum_{g \subseteq_G'} w_{gj} M_{gj}}{\sum_{g \subseteq_G'} w_{gj}}$$

where G is the total number of genes, g represents a single gene and j a single sample. N denotes the total read counts, and K_{gj} and K_{gr} denote the read counts for gene g and sample j and d_j is the scaling factor for the jth sample and r is the reference sample. Then $M_{gj} = log_2((K_{gj}/N_j)/(K_{gr}/N_r))$ and $w_{gj} = (N_j - K_{gj})/N_j K_{gj} + (N_r - K_{gr})/N_r K_{gr}$, where K_{gj} and $K_{gr} > 0$ [9].

2.2 Differential Analysis

Differential analysis in this study was performed with the edgeR package in R software. The chosen method to conduct the differential analysis was generalized linear model. GLMs use samples mean-variance relationship to specify probability distributions. The fitted log-linear model used is as follows

$$log\mu_{gi} = x_i^T \beta_g + logN_i$$

for each gene [11], [12]. Where, g represents each gene in the ith sample and x_i is a vector of covariates that specifies the treatment conditions applied to each sample i, and N_i is the total read counts in sample i. g is

a vector of regression coefficients by which the covariate effects are mediated for gene g.

For an experiment with multiple factors, like this one, the Cox-Reid adjusted likelihood method is used to estimate dispersions, also referred to as variance, in edgeR. Multiple factors were treated by using a design matrix when fitting the generalized linear model. Models were then fitted using the table of counts and its given design matrix. This method also accounts for all sources of variation which allows us to have a valid estimation of the dispersion. After the dispersion estimates were acquired and generalized linear models were fitted, the following design matrix was set up using model.matrix function, 0+day (day being E3, E8, etc.). The model was then fit using the generalized linear model approach, the R function glmFit was used. The model significance was tested using the likelihood ratio test corresponding R function: glmLRT.

Results were summarized with a list of top differentially expressed genes that were ranked from smallest to largest p-value including its gene information, log fold change (log-FC), average log counts per million (log-CPM), moderated test statistic, raw and adjusted p-value [13].

2.3 Multivariate Analysis

Multi-dimensional scaling (MDS) plots are used to explore the relationship between samples in all datasets. These plots are created using the average distance in all gene expressions between two samples. In short, samples will be plotted closer together if their gene expressions are similar to each other. Whereas samples that are plotted the farthest from each other are ones with the most different average gene expressions [12]. This plot is useful to investigate similarities between samples.

Heat maps are created by first performing two different hierarchal clustering analyses, one for the samples and one for the top 200 most variable genes. In hierarchal clustering the pairwise distance between two genes, or two samples, are calculated using the Euclidean distance. Genes or samples with the smallest pairwise difference are then clustered together. In heat maps color scaling is applied by row, or by gene, to represent the expression level of the gene in that sample. A color closer to red represents a gene that is highly expressed, while a color closer to blue represents a gene that is lowly expressed. This is useful for identifying changes in gene expressions across sample clusters.

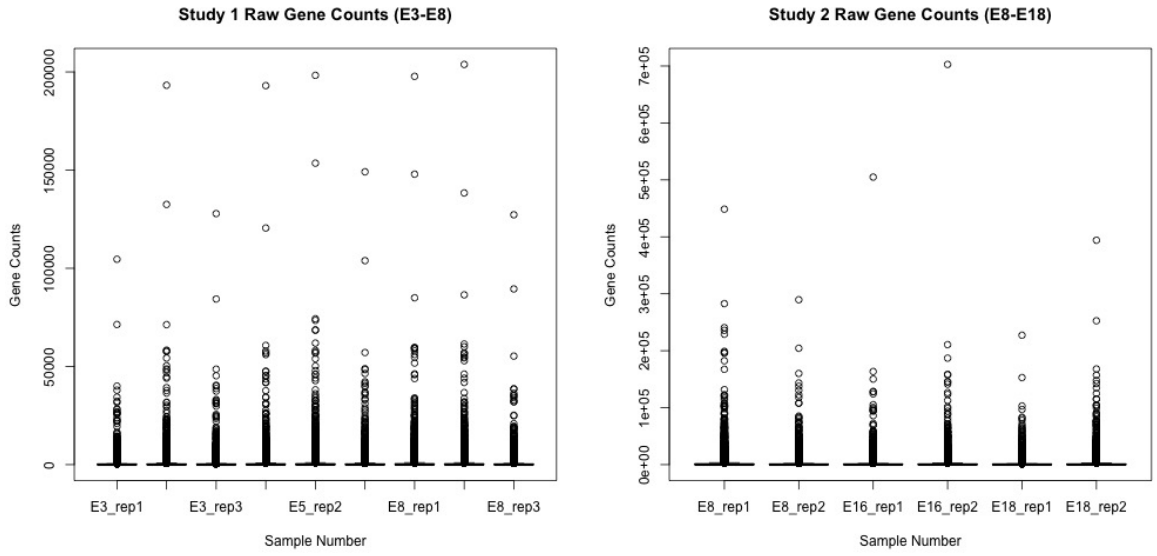
3 Results

3.1 Normalization

Prior to normalization, the raw genes counts in both studies are not suitable for a differential analysis because both distributions are highly right-skewed, see figure 1. As discussed early, TMM normalization was used

to adjust for the library size depth difference between samples. This is a crucial step because analyzing data with different library size depths will produce biased results. Data that is sequenced to a deeper level, or has a higher library size depth, will have a more precise quantification and identify a greater number of transcripts as compared to data that with smaller library depth size [14].

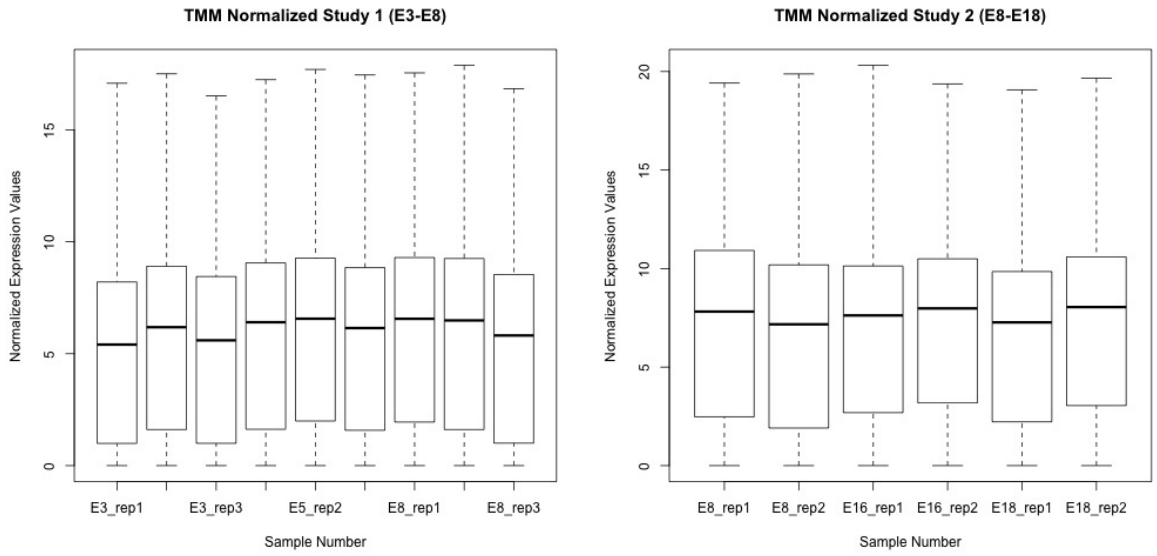
In figure 1, the distribution of raw gene counts for each sample prior normalization in shown for both studies. The distribution of raw gene counts in Study 1 have a smaller range compared to Study 2. After normalization, the distributions are still skewed but in a lot lesser degree, see figure 2.



(a) GSE89541

(b) GSE65938

Figure 1: Box plot of raw counts for Study 1 and 2



(a) GSE89541

(b) GSE65938

Figure 2: Box plot of TMM Normalized Counts for Study 1 and 2

3.2 Differential Analysis in Separate Studies

A total of 6 different differential analysis comparisons or tests were performed on the data sets. For study 1 these tests included E3 vs E5, E3 vs E8, and E5 vs E8. For study 2 these tests included E8 vs E16, E8 vs E18, and E16 vs E18.

The results of the differential analysis for each of the 6 comparisons are presented below in figure 3. For each comparison, the top 1000 genes are plotted using their –log10(PValue) versus log(FC).

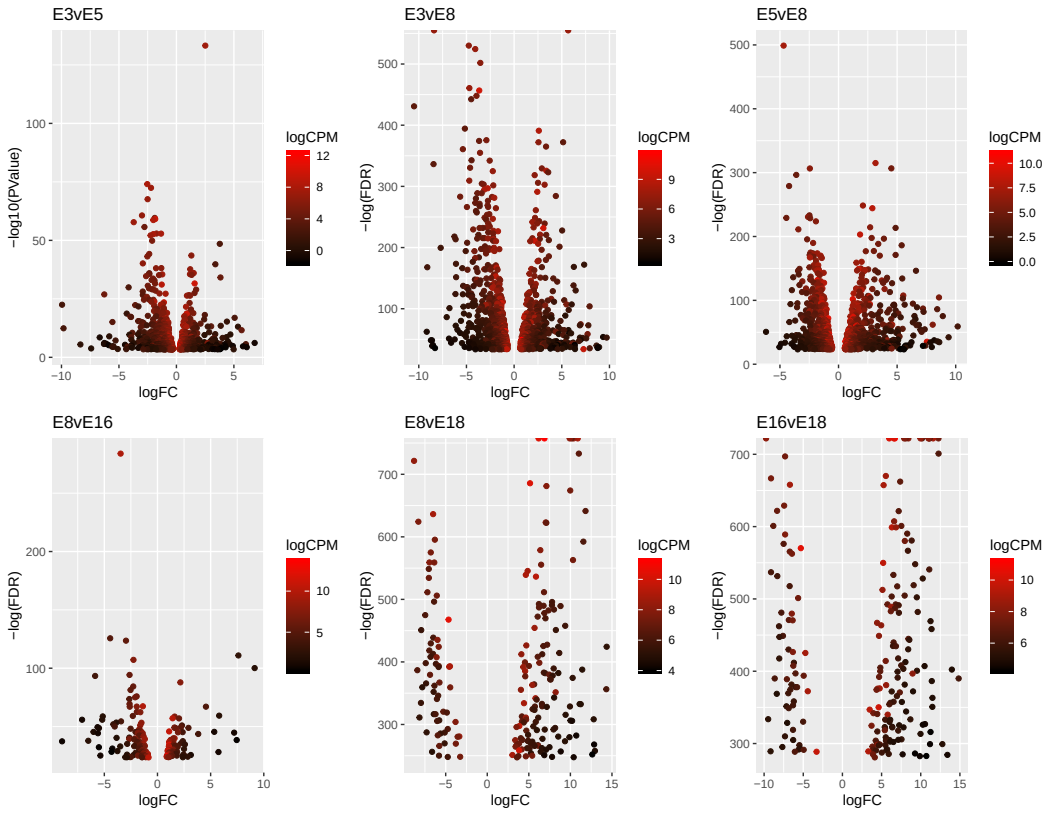


Figure 3: Volcano Plots for 6 Differential Analysis Tests

Although it is interesting to observe genes that are significantly expressed between two time points of development, our main interest was to identify a set of key genes that were significantly expressed at multiple points in the developmental process. To find this the top 200 significantly expressed genes from each test were extracted and used for further analysis and comparison. First, the number of 'alike' genes that came up in exactly 2 tests were counted and recorded. This comparison was done within each study. See tables 1-2 for the count breakdown in both studies.

	E3vE5	E3vE8	E5vE8
E3vE5	200		
E3vE8	107	200	
E5vE8	66	136	200

Table 1: Count of Significantly Expressed Genes in Both Tests in Study 1

	E8vE16	E8vE18	E16vE18
E8vE16	200		
E8vE18	169	200	
E16vE18	17	28	200

Table 2: Count of Significantly Expressed Genes in Both Tests in Study 2

After this was done, all 6 tests were then used to see how many genes were significantly expressed in 3 or more tests. Our results showed that 77 genes were found to be significantly expressed in three tests, 3 genes in four tests, and 3 genes in 5 tests. There were no genes found to be significantly expressed in all 6 tests. The list of genes that were significantly expressed in 4 and 5 tests are listed below. These 6 genes became a set of interest for us.

List of Genes Significantly Expressed in 5 out of the 6 Tests

1. ENSGALG00000033304
2. ENSGALG00000003045
3. ENSGALG00000002375

List of Genes Significantly Expressed in 4 out of the 6 Tests

1. ENSGALG00000013956
2. ENSGALG00000003582
3. ENSGALG00000038515

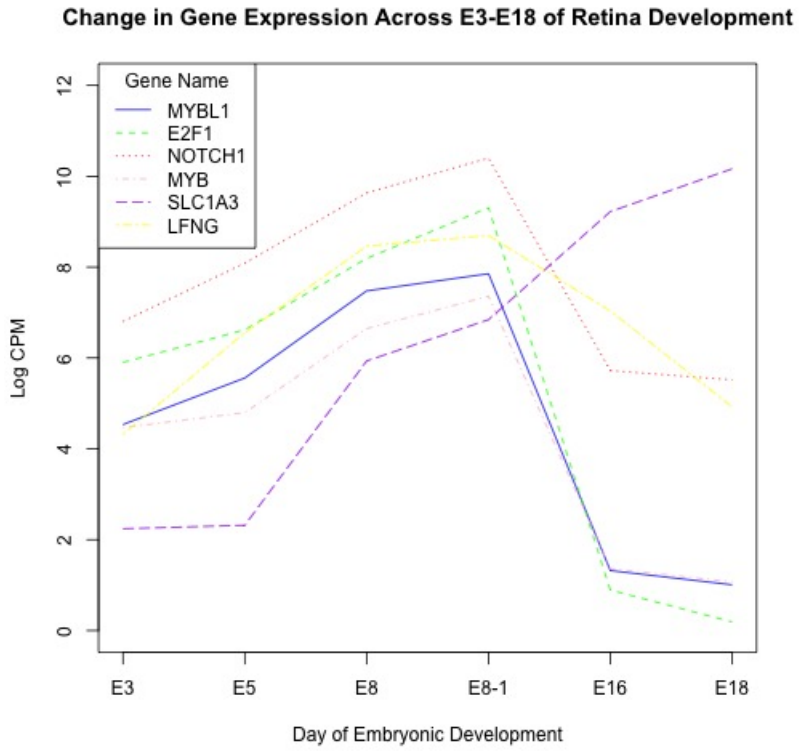


Figure 4: Gene expression patterns of selected 6 genes of interest across all days of development

Figure 4 shows the gene expression patterns of the 6 genes of interest across days 3 to 18 of development. Genes MYB1, E2F1, NOTCH1, MYB, and LFNG all have similar patterns. Their expressions rise from days 3 and 5 and peak at day 8 of embryonic development. Following day 8, their expressions become much lower. On the contrary, the SLC1A3 gene expression levels continue to rise between each day of development.

3.3 Differential Analysis Across Total Time-frame of Development

After the separate differential analysis tests were run for each study, the datasets were then combined together to allow for a test from day E3 to day E18 of retina development. The results of this test gives us an insight to which genes are most significant from the beginning to the end of chicken retina's development. See figure below for volcano plot created using the top 1000 genes from the test.

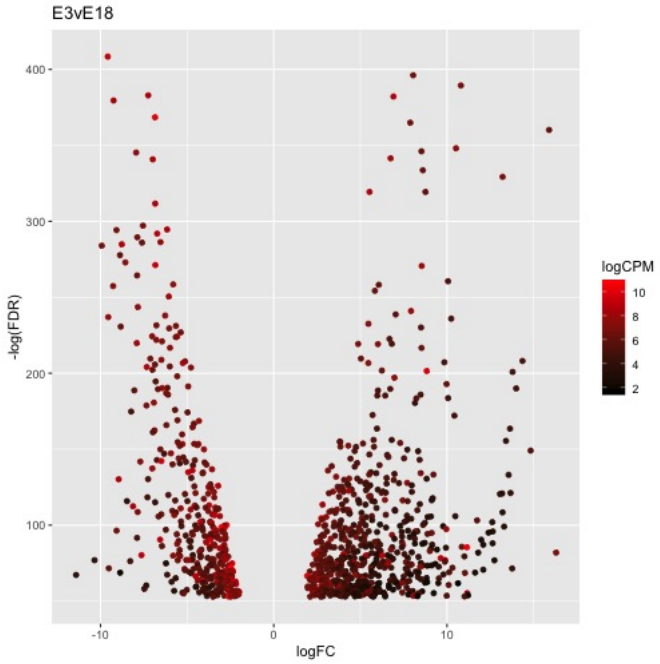


Figure 5: Volcano Plot for Results from E3 vs E18 Test

In the table below a summary of test results for the top 5 most significant genes between day 3 and 18 are highlighted. Figure 6 shows the gene expression patterns for each of the 5 genes at each time point of development. All 5 genes appear to have little to no change in expression until after day 8 of development. The expression levels of the CDK1 and MOB3B genes lower after day 8, while the expression levels of the PDE6C, CLUL1, and ENO2 genes all rise.

Gene Name	logFC	logCPM	LR	PValue	FDR
MOB3B	-9.5835	9.3283	829.8824	1.72E-182	4.28E-178
PDE6C	8.0448	6.6259	803.9733	7.38E-177	9.18E-173
CLUL1	10.8007	6.7601	789.7776	9.01E-174	7.47E-170
CDK1	-7.2482	8.6068	776.1236	8.38E-171	5.21E-167
ENO2	6.9082	8.8819	774.2566	2.13E-170	1.06E-166

Table 3: Top 5 Significantly Expressed Genes in E3vE18 Test

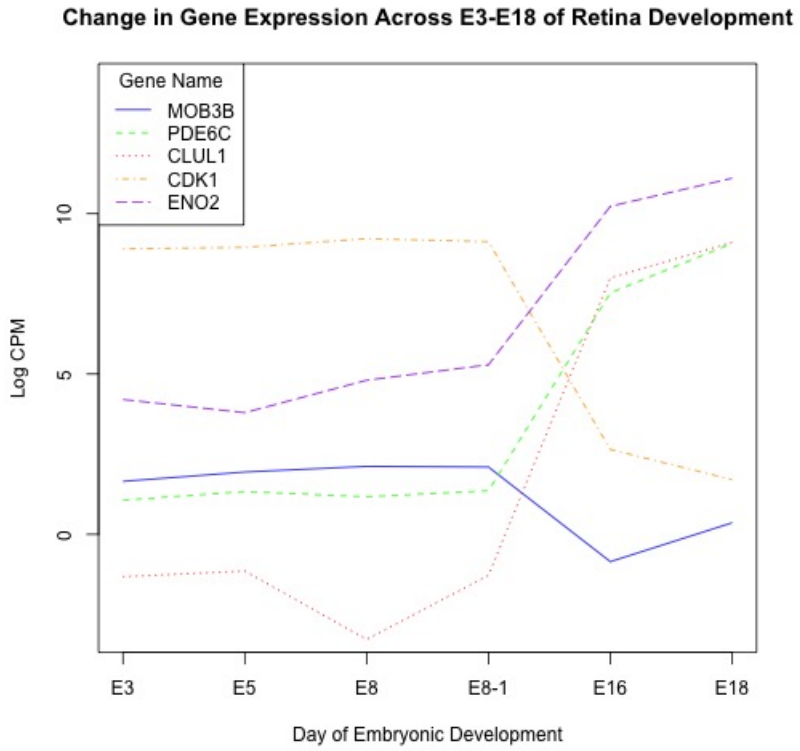


Figure 6: Gene expression patterns of top 5 significantly expressed genes from E3vE18 test across all days of development

3.4 Multivariate Analysis

In this part of the analysis we use different multivariate approaches to observe the behavior of samples and genes. In figure 4, a multi-dimensional scaling (MDS) plot is used to show the average expression difference between all genes in each sample. Each sample is labeled and colored based on the group day it belongs too. The groups are colored as follows: E3 green, E5 blue, E8 light blue, E16 black, and E18 red. Note that E8rep1.1 and E8rep1.2 are day 8 samples from study2.

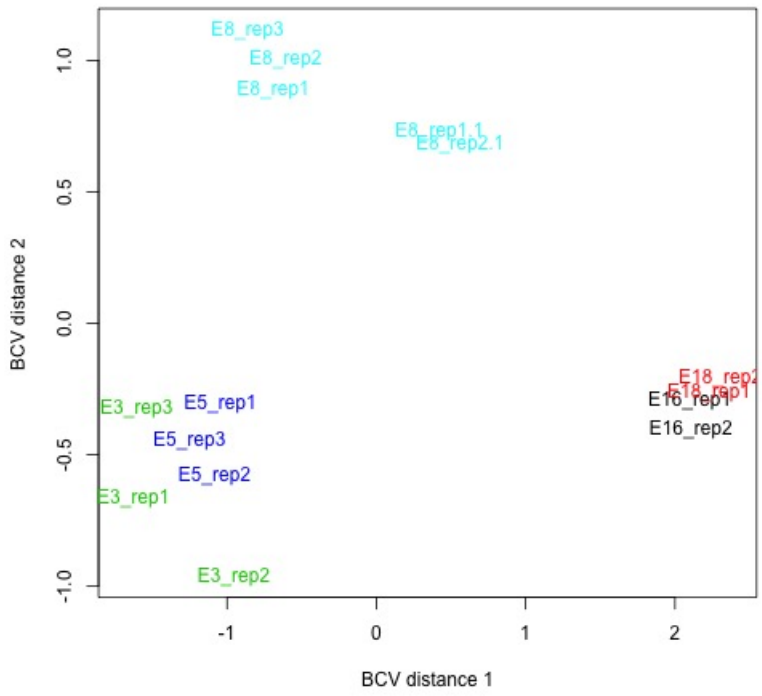


Figure 7: MDS plots showing average gene expression differences between samples

Here we can observe that samples from day 3 and 5 tend to be closely related as well as samples from day 16 and 18 of development. We also see that all samples from day 8 appear to be closely related on the graph. This confirms that our normalization method was effective because day 8 of development was the only day of sample overlap between the two studies.

To further investigate sample clustering, a hierarchical heat map was created using the top 200 most variable genes in the dataset, see figures 5-6. Again, different colors were assigned to distinguish between sample groups in both data sets. In study 1 the groups are colored as follows: E3 purple, E5 orange, and E8 blue. In study 2 the groups are colored as follows: E8 blue, E16 purple, and E18 orange. The colors inside the histogram can be interpreted using the color key in the upper left corner. A color shade closer to dark red represents genes that are highly expressed in the group and color shade closer to dark blue represent genes that are lowly expressed in the group.

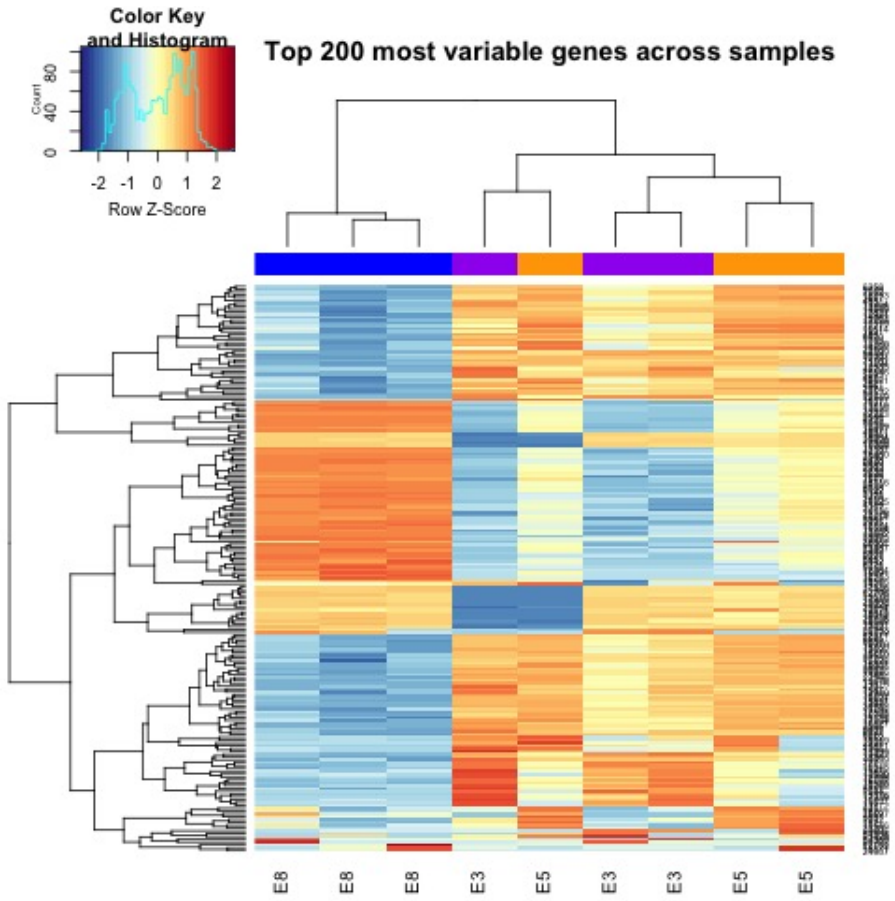


Figure 8: Study1 Hierarchical Clustering Heatmap

In figure 8, samples from day 8 cluster nicely together, but there seems to be some overlap in samples from day 3 and 5. This clustering unfortunately makes the heat map interpretation not that helpful.

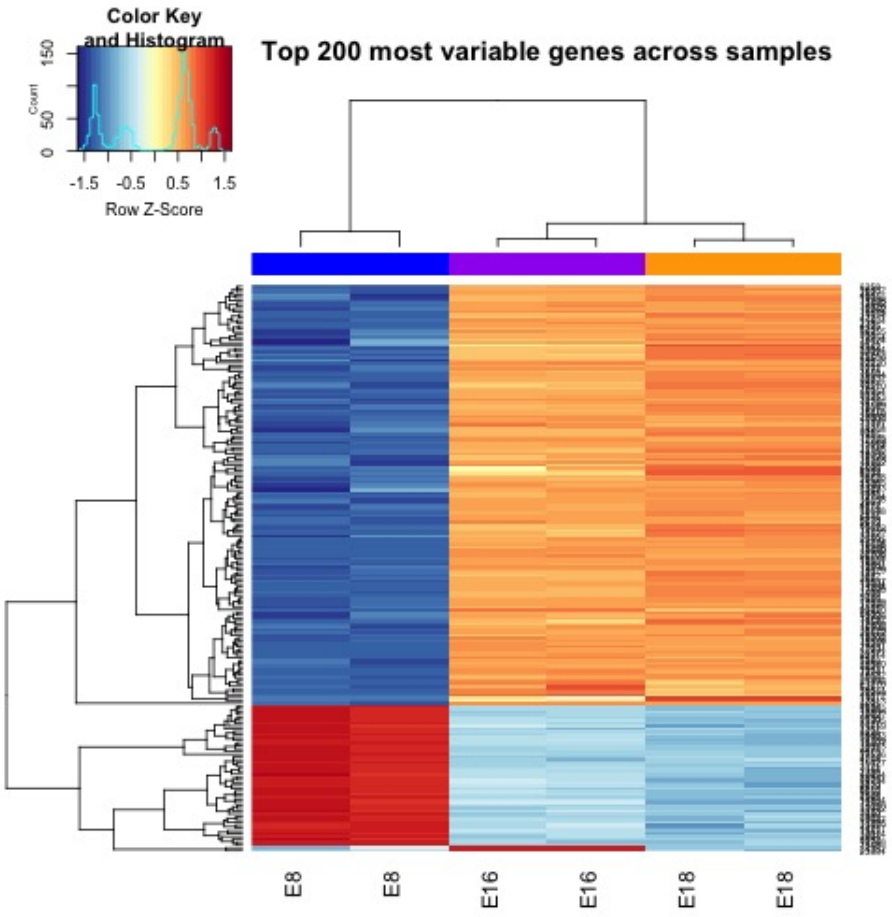


Figure 9: Study2 Hierarchical Clustering Heatmap

In figure 9, study 2 samples clustered into 3 distinct groups based on the sample days. This is expected based on the results in MDS plot. This heat map also shows a clear difference in gene expressions with day 8 compared to day 16 and 18. Genes that are lowly expressed in day 8 show up as being highly expressed in day 16 and 18 (color change from dark blue to orange). Genes that are highly expressed in day 8 show up as being lowly expressed in day 16 and 18 (color change from red to blue). There appears to be no major change between day 16 and 18. The genes only become slightly higher or lower expressed at day 18. (colors shade becomes darker).

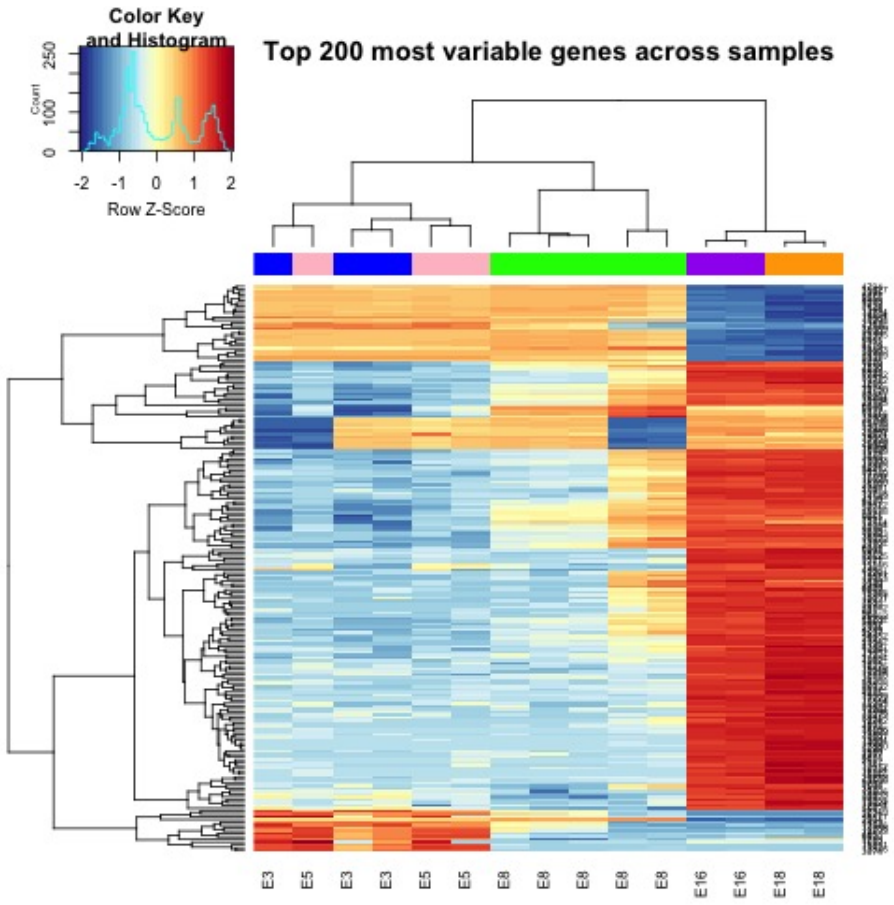


Figure 10: Studies Combined Hierarchical Clustering Heatmap

This heat map based on combined data reveals 3 major clusters for samples. Day 3 and 5 from one cluster, day 8 from both studies form the second group, and the final group is day 16 and 18. In terms of gene expression values, there are two general clusters. Genes have similar expression values across samples from Day 3, 5 and 8 (both studies). Samples from day 16 and 18 form the second cluster with similar gene expression values. Most of the genes have low to moderate expression values across E3, E5, and E8 compared to E16 and E18.

4 Conclusion

Our main goal in this study was to use an effective normalization method to accurately compare and combine data from two alike RNA-seq studies. In this study we decided to use the trimmed mean of M-values method (TMM). The results from the MDS plot confirm that the TMM normalization was successful in accounting for the differences in library size depth. This plot shows samples from day 8, in both experiments, clustered

together, confirming that the differences between the two samples was minor and our normalization method did in fact work.

Being able to combine the data from these two studies allowed us to study the chicken retina developmental process across days 3-18 of development. Throughout this analysis we discovered a few notable patterns in the data. First, it appears that day 3 and day 5 of development share similar patterns. These days were clustered closely together on both the MDS plot and clustering heat maps. This observation is again confirmed when studying the gene expression pattern graphs for a select number of genes. The change in each gene's expression between day 3 and day 5 show little to no change. These same results were observed between day 16 and day 18 of development. On the other hand, day 8 appears to be an important day in the developmental process. The samples from day 8 are separated from the rest of the days of development in all clustering figures. Even on the gene expression graphs day 8 seems to be turning point in expression level for a given gene. Most of the genes that we highlighted show that their expression levels either drop off significantly following day 8 or continue to increase.

All together the results from our gene expression analysis on the normalized data helped us identify a set of genes that could possibly affect the retina development process. This analysis also provides a framework for other researchers, interested in studying retina development, to use to study genes of their interest as well.

References

[1] Rasoul, Bejan Abbas, "Characterizing Epigenetic Regulation in the Developing Chicken Retina" (2018). *Masters Theses*. 569.

[2] Langouet-Astrie, Christophe J., et al. "RNA Sequencing Analysis of the Developing Chicken Retina." *Scientific Data*, vol. 3, 2016, p. 160117., doi:10.1038/sdata.2016.117.

[3] "Retinal Disorders — Retina — Macular Degeneration — MedlinePlus." *MedlinePlus Trusted Health Information for You*, medlineplus.gov/retinaldisorders.html.

[4] Home - GEO - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/geo/

[5] Law, Charity, et al. "RNA-Seq Analysis Is Easy as 1-2-3 with Limma, Glimma and EdgeR." *Bioconductor - Open Source Software for Bioinformatics*, 21 Sept. 2017, www.bioconductor.org/help/workflows/RNAseq123/.

[6] Kim, Daehwan, et al. "TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions." *Genome Biology*, vol.14, no.4, 2013, doi:10.1186/gb-2013-14-4-r36

[7] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R. Gingeras; STAR: ultrafast universal RNA-seq aligner, *Bioinformatics*, Volume 29, Issue 1, 1 January 2013, Pages 15–21

[8] Li, Peipei, et al. "Comparing the Normalization Methods for the Differential Analysis of Illumina High-Throughput RNA-Seq Data." *BMC Bioinformatics*, vol. 16, no. 1, 2015, doi:10.1186/s12859-015-0778-7.

[9] J. Zyprych-Walczak, A. Szabelska, L. Handschuh, et al., "The Impact of Normalization Methods on RNA-Seq Data Analysis," *BioMed Research International*, vol. 2015, Article ID 621690, 10 pages, 2015. doi:10.1155/2015/621690

[10] Rodrigues, Tania, et al. "Delayed Neurogenesis with Respect to Eye Growth Shapes the Pigeon Retina for High Visual Acuity." *Development*, vol. 143, no. 24, 2016, pp. 4701–4712., doi:10.1242/dev.138719.

[11] Lu, J., Tomfohr, J., and Kepler, T. Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics* 6, 165, 2005

[12] Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 18, 11–94, 2010

[13] Robinson, M. D., et al. "EdgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics*, vol. 26, no. 1, 2009, pp. 139–140., doi:10.1093/bioinformatics/btp616.

[14] Conesa Ana, et.al. "A survey of best practices for RNA-seq data analysis" *Genome Biology*, Volume 17, 2016.

[15] Merchant, Nirav, et al., "The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences," PLOS Biology (2016), doi: 10.1371/journal.pbio.1002342.

[16] (n.d.). Retrieved from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[17] Yang Liao, Gordon K. Smyth, Wei Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, *Bioinformatics*, Volume 30, Issue 7, 1 April 2014, Pages 923–930, https://doi.org/10.1093/bioinformatics/btt656