ROLE OF *Top2b* IN PHOTORECEPTOR GENE REGULATORY NETWORK BY SINGLE-CELL TRANSCRIPTOME ANALYSIS

BY

ALEXANDRIA PINTO

A thesis submitted to the
School of Graduate Studies
Rutgers, The State University of New Jersey

In partial fulfillment of the requirements
For the degree of
Master of Science
Graduate Program in Biomedical Engineering
Written under the direction of
Li Cai
And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2018

ABSTRACT OF THE THESIS

ROLE OF *Top2b* IN PHOTORECEPTOR GENE REGULATORY NETWORK BY SINGLE-CELL TRANSCRIPTOME ANALYSIS

By ALEXANDRIA PINTO

Thesis Director:

Dr. Li Cai

TOP2B is an enzyme that allows for access to the DNA strand for gene transcription. During development, TOP2B is found in cells which have finished mitosis and proliferation, suggesting its function in cell differentiation. Previously, bulk RNA-seq analysis of the retina revealed TOP2B controls expression of genes in the photoreceptor gene-regulatory network. However, bulk RNA-seq does not allow for direct analysis of individual cells to identify the role of TOP2B in photoreceptor cell differentiation. The central hypothesis is that grouping cells based on the photoreceptor gene regulatory network and applying bioinformatics analysis to the data can show that TOP2B plays an essential role in proper photoreceptor differentiation. In this study, we preform bioinformatics analysis on publically available single-cell RNA-seq (scRNA-seq) dataset of postnatal day 14 mouse retina (GSE63473) to determine to role of TOP2B in the photoreceptor gene regulatory network and identify novel genes which contribute to this pathway. Analysis of photoreceptor scRNA-seq data reveals that TOP2B expression is

correlated with the expression of photoreceptor marker genes, confirming its role in photoreceptor differentiation. In addition, gene *Fam19a3* was identified for its novel role contributing to the Top2b-controlled photoreceptor gene regulatory network. Thus, this study provides further insight into the photoreceptor differentiation processes that could be affected by the gene regulatory pathway.

Acknowledgements

I would like to thank:

- My advisor, Dr. Li Cai for his support and encouragement throughout this project. His insight and influence helped make my project great.
- Dr. Jay Sy and Dr. Ilker Hacihaliloglu for agreeing to be on my committee and a special thanks to Dr. Adam Gormley for stepping in on short notice.
- My friends for continually pushing me to do better and supporting me when I felt overwhelmed.
- My parents for constantly believing in me.
- My boyfriend for being my best friend and sounding board.
- My lab members for their understanding of my project and input
- The members of Stack Overflow and Bioconductor's help page, for answering my questions when I could not.
- My cat for being serene and calming me when I felt frustrated.

Dedications

I would like to dedicate my thesis to:

- My parents for supporting me through my education.
- Dan Kirlin for living with me and being with me through this whole process. I would not have been able to do this without you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Single cell RNA-sequencing (scRNA-seq) is becoming a popular data analysis technique in biomedical research. scRNA-seq is capable of comparing individual cell transcriptomes, allowing for categorization of similarities and differences in a population of cells, identification of rare cell types and trace lineage and development of cells [1]. While conceptually similar to bulk RNA-seq analyses, scRNA-seq tends to have a smaller number of gene counts detected and needs to be normalized by counts per million mapped reads, have a non-normally distributed expression measurement, and reveal higher biologic variability [1]. After normalization, typical statistical methods can be performed, such as the students t-test which has been proven to work on log normalized data[23]. scRNA-seq data requires filtering techniques to ensure the best cells are used for analysis, and at the moment there does not exist a gold standard [1]. Studies utilizing scRNA-seq have provided discoveries on the cellular level since the first published study in 2009, such as classification of retinal bipolar neurons [2], dissection of cell types in tissues [3], and identify the variety of cells in the cortex and hippocampus [4].

In 2015, the article "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets" by E. Macosko was published to *Cell* describing a new method of obtaining single cell RNA-seq data, Drop-seq. This method creates nanoliter droplets holding individual cells that are barcoded for identificationThe dataset used in this thesis is obtained from 3T3 mouse retina cells from Macosko's article using the Drop-seq method. Seven mice retinas were digested in a papain solution and the tissue was titrated to generate a single cell suspension. The suspended cells then were

used in the Drop-seq method, which is the primary method described in the paper. The single cell suspensions and barcoded beads are joined in a microfluidic device with oil pumping through it. Flow rates differ for each suspension to allow for a single cell and single barcoded primer to be in each droplet. Droplets are broken and centrifuged in the same tube to obtain the mRNA in all the cells and then reverse transcribed. The resulting sequenced library will have all the mRNA of the tissue in one dataset, that is then sorted into a count matrix containing the barcoded cells as columns and the genes as rows. In each entry is the number of transcript reads, or "counts," of each gene for each cell[6].

Drop-seq is currently one of the best methods for obtaining scRNA-seq libraries. Since scRNA-seq is a relatively new method, Drop-seq has some room for improvement. The droplets are vulnerable to impurities and using the microfluidic device does not guarantee that only one cell will be present in each droplet. Macosko et al also claimed that the retina dataset that we use in this analysis has about 10% of all cells in the tissue being either doublets or impure. However, this is the only publically available scRNA-seq retina dataset.

The retina has become an area of interest in single cell RNA sequencing analysis. The tissue contains a diverse amount of cell types and has also proven to have additional subtypes. Single cell RNA sequencing has been useful in defining subsets of retinal ganglion cells [8] and retinal bipolar cells [2]. However, this method has not yet been utilized on the photoreceptor cells. Photoreceptor cells are the most abundant cell type accounting for >65% [6] of the retinal cells, with rods composing 97.2% [9] of photoreceptors. For this reason, our focus is primarily rods, with some interest in cone expression. Rods communicate with the bipolar cells during phototransduction and

Shekhar et al. [2] found that there exist fifteen types of bipolar cells in the retina. Because of this, we hypothesize that there exist subtypes of rods corresponding to the types of bipolar cells.

The first step in identifying rod subtypes is identifying rods through rod marker genes. Rod marker genes primarily include genes involved in phototransduction, such as *Rho, Pde6b* and *Pde6a.* Marker genes and their functions are shown in **Table 0.** *Nrl* and *Crx* are proven to regulate phototransduction genes and photoreceptor differentiation [10]. Below is a simplified map of the gene regulatory networks with transcription factors at the top of the map and contributing to changes in target genes **(Figure 1)** [10].

Gene	Marker For:	Function
Rho	Rods	Rhodoposin coding gene, essential for vision in low-light conditions
Sag	Rods	Inhibits rhodopsin from coupling and preventing transductions
Rcvrn	Rods	Regulates rhodopsin kinase
Pde6g	Rods	Functions in the phototransduction signally cascade
Opn1sw	Cones	G-protein coupled receptor that allows for blue-yellow/short wavelength sensory
Opn1mw	Cones	G-protein coupled receptor that allows for green/medium and long wavelength sensory in the cone photoreceptors.
Gnat2	Cones	Creates the cone-specific alpha subunit of transducin in phototransduction

Table 0: Marker genes and their functions

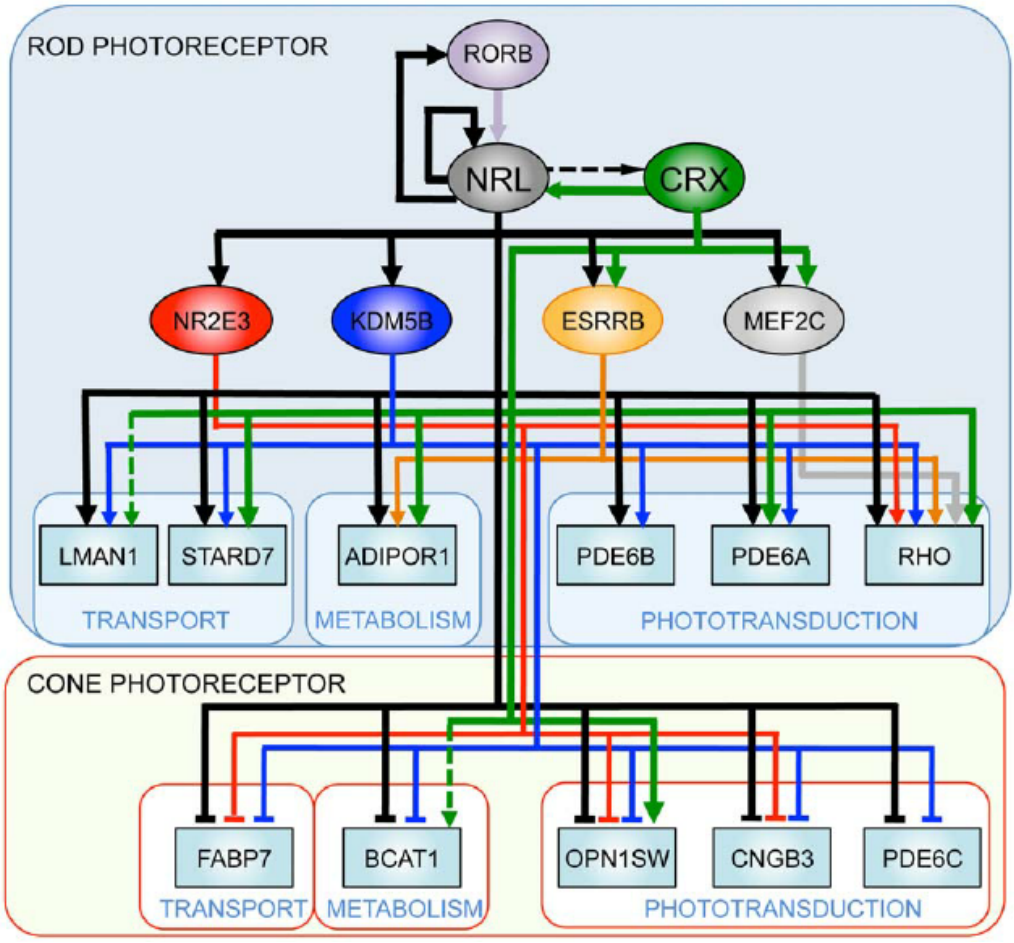

Figure 1: Simplified gene regulatory network in rod and cone photoreceptors from H. Hao et al's paper "Transcriptional Regulation of Rod Photoreceptor Homeostasis by *In Vivo* NRL Targetome Analysis." **[10]**

Topoisomerase IIβ (TOP2B) is an enzyme that changes and controls the topological states of double stranded DNA[7]. *Top2b* is expressed in progenitor cells that have gone through the final division and are in neural development [5]. Previously, we found that *Top2b* controls expression of key genes in the photoreceptor gene-regulatory network (*Crx, Nr2e3, Opn1sw*, and *Vsx2*) **(Figure 2)** and has a role in late stage photoreceptor differentiation and maturation using bulk RNA-seq [5]. Deficiency of *Top2b* in retinal cells has been found to cause defects in the retina. Here we utilize single cell RNA sequencing analysis to identify photoreceptors, possible subtypes of

photoreceptors and identify genes of interest that may contribute to *Top2b*'s role in the photoreceptor gene regulatory network. (GSE63473)[6].

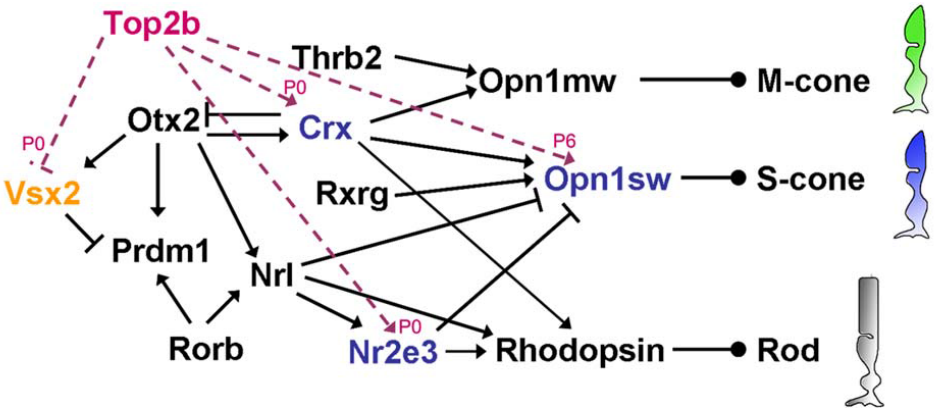

Figure 2: Schematic drawing of photoreceptor transcriptional network with Top2b as key modulator.[5]

Chapter 2

Methods

We acquired a dataset that has been publicly shared on NCBI by the McCarroll Lab from Harvard University originally part of E. Macosko et al article "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets"[6]. From this dataset, we downloaded the P14 retina merged digital expression matrix. The data was converted into R and reformatted so it could be read into the scater package[11]. Scater compressed the data into a more usable form and calculated quality control metrics, which includes total gene transcript counts in each cell, total counts for each gene, and the average expression of each gene. We kept cells with a high gene transcript count expression, setting 4000 as the lowest number of counts a cell could have. This resulted in 2334 cells kept of the 49300 cells.

Cells were then broken into categories, TOP2B negative or positive, NRL negative or positive, and RHO negative or positive. Using the same dataset, the cells were broken down once again by TOP2B negative or positive, CRX negative or positive and RHO negative or positive. Cells were deemed negative if there is not a single transcript for the gene of interest being expressed in the cell. Cells are positive if at least one transcript of the gene is found. A total of 16 groups were found and of those groups four contained the genes that are part of the photoreceptor gene regulatory network. **(Figure 3, Figure 4)**. These four groups of cells: *Top2b+Nrl+Rho+, Top2b+Crx+Rho+, Top2b-Nrl(+)Rho(+) and Top2b(-)Crx(+)Rho(+)* cells, denoted respectively as T+N+R+, T+C+R+, T-N+R+ and T-C+R+ cells respectively from here on), are used for analysis.

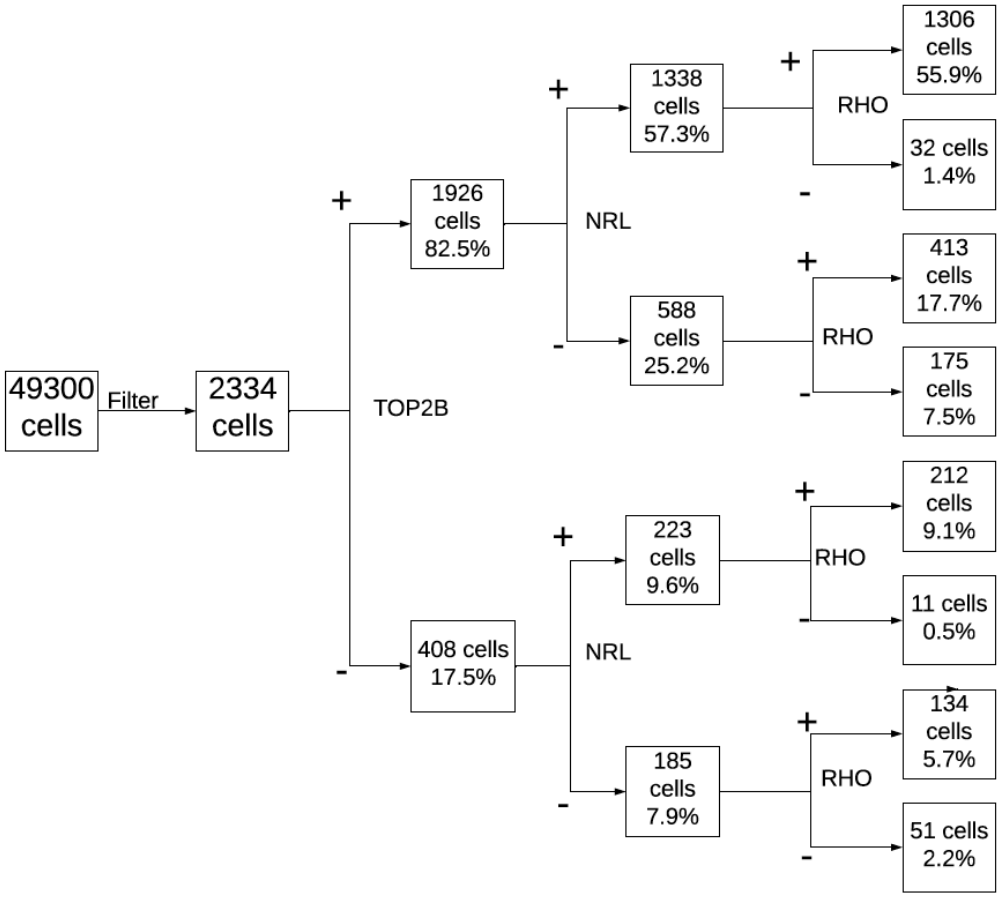

Figure 3: A tree showing the breakdown of GSE63473 dataset based on NRL. The filter breakpoint is based on the total features being expressed in each cell. Cells with more than 4000 features are kept for further analysis. Branch points were created based on the presence or absence of a single transcript of TOP2B, NRL and RHO. The right most groups are then used for further analysis.

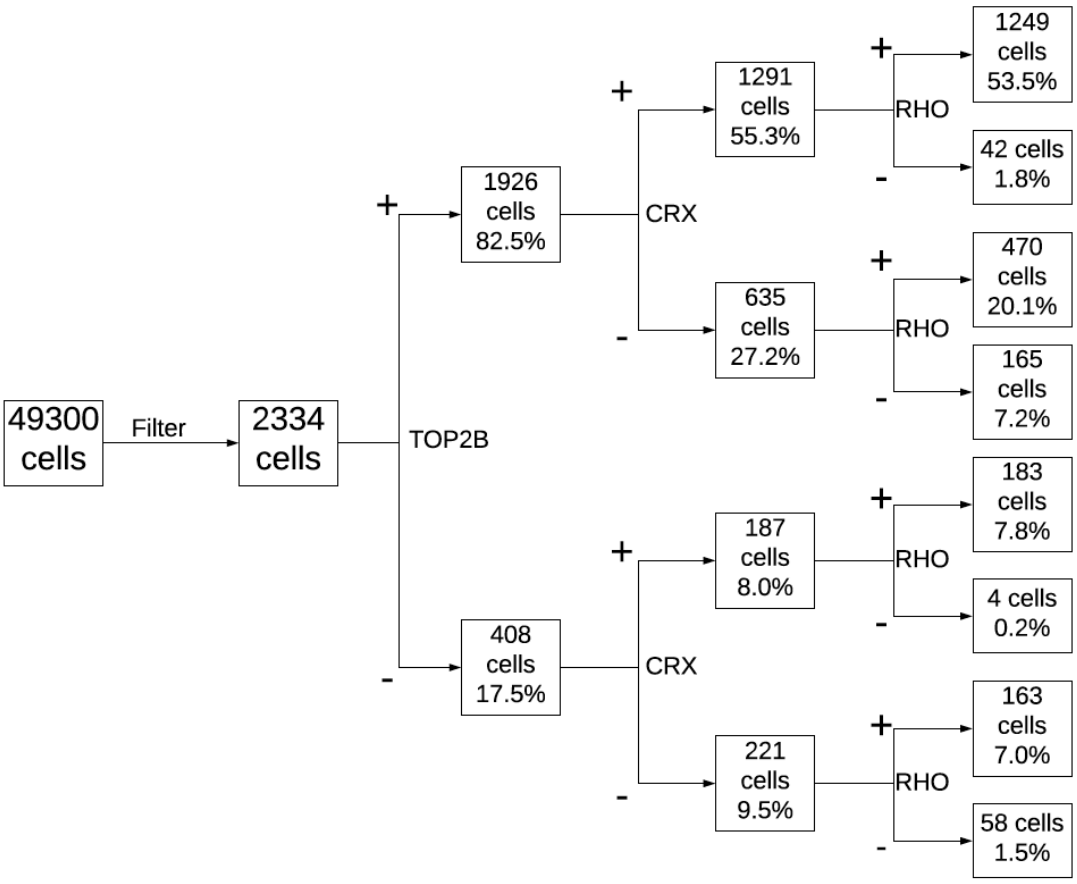

Figure 4: A tree showing the breakdown and cell counts of GSE63473 dataset based on CRX. The filter breakpoint is based on the total features being expressed in each cell. Cells with more than 4000 features are kept for further analysis. Branch points were created based on the presence or absence of a single transcript of TOP2B, CRX and RHO. The right most groups are then used for further analysis.

All groups underwent the same analysis to identify differentially expressed genes. The method was adapted from E. Macosko et al article "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets"[1]. Counts were converted into counts per million and log transformed. Genes that were not expressed in any of the cells were removed. Variance, or the measure of how the log counts of each gene is dispersed around the mean of the gene, was calculated using the variance function in R. Dispersion measure was also calculated dividing the variance of each gene by its mean log counts. Genes were than separated into similarity bins based on mean log expression of each gene. A z-score was then found for all genes, genes with less than a 1.7 z-score were removed. The z-score for each gene is calculated by subtracting the average dispersion measure of all genes in a single bin from the dispersion measure of the gene and divided by the standard deviation of the dispersion measure across all genes in the bin. This process is meant to identify differentially expressed genes that have a high variation of expression across all cells. These genes can then be used to identify significant differences in cells, potentially identifying different subtypes of cells. Genes of interest were kept for analysis as well.

Finally, SC3[12] was used for unsupervised clustering of each group. SC3 identifies clusters and confidence measures describing how correlated the cells are within a cluster. This method has been proven to work well with scater and is also the newest and simplest method of clustering single cell RNA-seq data at the moment.

For additional analysis, a comparison algorithm was created. This allows us to compare the top 100 highest expressed genes between groups, identifying genes that are shared in groups and genes that are only present in one. The algorithm utilizes scater's

ranking quality control metric generated automatically. This is the easiest method of getting a summary of the single cell RNA-seq data at the moment. The method generates a breakdown of the data including the total number of different genes expressed in each cell, the total number of cells a gene is found in and rank of each gene. The ranking system labels genes in such a way that the gene that is most highly expressed in the most number of cells has the rank of 1 and a gene that is lowly expressed in only one cell is last. Excel sheets were generated to summarize gene expression. To compare differences in gene expression between groups, R's built in function of the student's t test is used on the log counts of each gene.

Chapter 3

Results

3.1 Typical rod marker genes expressed in T+N+R+ and T+C+R+ cells

I wanted to determine which gene regulatory network more accurately identifies rods and determine if any genes are contributing significantly to either network. Looking back at **figures 3 and 4**, *Top2b* is expressed in a majority of all cells at 82.5%. In both the *Nrl* and *Crx* branching, more than half of all cells are in the T+N+R+ and T+C+R+ groups, with 2% more cells in the T+N+R+ group.

Since the T+N+R+ and T+C+R+ groups are formed from the same dataset, I assumed a majority of the cells are shared between groups. There are 1306 cells in the T+N+R+ group and 1249 cells in the T+C+R+ group, giving a difference of 57 cells. To confirm that these cells are the same, I isolated cells in the T+N+R+ so the gene expression of the isolated cells is T+N+C-R+ and in the T+C+R+ so the gene expression of the remaining cells is T+N-C+R+. In the isolated T+N+C-R+ group, there are 175 cells that do not express CRX; while in the isolated T+N-C+R+ group there are 118 cells that do not express NRL. The difference in isolated cell groups is also 57. This implies that the cells in the all positive group are all the same cells except for the 175 cells in the T+N+C-R group and the 118 cells in the T+N-C+R+ group. To confirm this, I compared the cell names of both T+N+R+ and T+C+R+ groups and found this assumption to be correct.

3.1.1 Gene expression in T+N+R+ and T+C+R+ cells are nearly identical

Of the top 100 genes being expressed in each group, 92 are shared between the T+N+R+ group and the T+C+R+ group. The genes that are shared are mostly marker genes for photoreceptors, with the top expressed gene in both groups being *Rho*, the gene creating rhodopsin (see appendix for all 100 genes). Other genes include *Sag,* a gene whose protein inhibits rhodopsin from coupling and preventing transductions, *Pde6g*, which functions in the phototransduction signally cascade, and *Rcvrn* which helps regulate rhodopsin kinase. The large expression of these genes indicates that the cells in these groups are photoreceptor cells. This means that the network including *Top2b* → *Nrl/Crx* → *Rho* is a good indicator of photoreceptor differentiation. The expressions of these known photoreceptor marker genes are very similar in both groups, as indicated by the p-values obtained from the student's t-test listed in **table 1**. P-values indicate that differences between most of the marker genes are negligible. This is likely due to the larger number of cells that are shared between groups. However, the expression of *Crx* has a significant difference with a p-value of 4.7e-9 while the *Nrl* expression is insignificant. Since the cells are largely the same, the difference likely derives from the cells that they do not have in common. This will be looked into further when the groups are separated in section 3.1.2.

Additional analysis was done on cone marker genes, including *Opn1sw*, *Opn1mw* and *Gnat2*. *Opn1sw* is a g-protein coupled receptor that allows for blue-yellow/short wavelength sensory in the cone photoreceptors and *Opn1mw* is a g-protein coupled receptor that allows for green/medium and long wavelength sensory in the cone photoreceptors. Lastly, *Gnat2* creates the cone-specific alpha subunit of transducin in

phototransduction[14]. This analysis is meant to identify whether or not *Crx* has a higher chance of creating cone photoreceptors. The p-values found for the cone specific genes indicate that no significant difference exists between the T+N+R+ and T+C+R+ groups.

Table 1: Highly Expressed Genes in Both TOP2B(+)NRL(+)RHO(+) And TOP2B(+)CRX(+)RHO(+) Cells on a logarithmic scale

		Maximum	Mean	Minimum	Expressed in _ Cells	Variance	Dispersion Measure	z Score	p value
RHO									
	NRL(+)	7.7533	5.3973	0.4639	1306	2.6921	0.4988	-0.5870	
	CRX(+)	7.7597	5.4097	0.5582	1249	2.8676	0.5301	-0.4252	
									0.8513
TOP2B									
	CRX(+)	4.1993	2.0461	0.5079	1249	0.4403	0.2152	-1.5879	
	NRL(+)	4.1932	2.0297	0.4499	1306	0.4335	0.2136	-1.7027	
									0.5306
NRL									
	CRX(+)	5.7207	3.1270	0.0000	1131	2.3012	0.7359	0.3348	
	NRL(+)	5.7144	3.2130	0.2943	1306	1.6876	0.5252	-0.4834	
									0.1247
CRX									
	CRX(+)	4.5792	2.6009	0.2778	1249	0.9077	0.3490	-1.0939	
	NRL(+)	4.5730	2.3431	0.0000	1131	1.5632	0.6672	0.0719	
									0.0000
SAG									
	CRX(+)	7.5714	5.1430	0.0000	1216	3.1482	0.6121	-0.1223	
	NRL(+)	7.5650	5.1152	0.0000	1275	3.0033	0.5871	-0.2413	
									0.6891
PDE6G									
	CRX(+)	7.2377	4.1879	0.0000	1181	3.2310	0.7715	0.4662	
	NRL(+)	7.2312	4.1431	0.0000	1241	3.1658	0.7641	0.4512	
									0.5272
RCVRN									
	CRX(+)	6.4911	3.7887	0.0000	1148	3.0952	0.8170	0.6341	
	NRL(+)	6.4847	3.7816	0.0000	1226	2.8626	0.7570	0.4232	
									0.9182
OPN1SW									
	CRX(+)	7.0401	1.0478	0.0000	355	3.5076	3.3477	9.9784	
	NRL(+)	7.0337	0.9980	0.0000	354	3.3897	3.3965	11.3866	
									0.4983
OPN1MW									
	CRX(+)	5.5870	0.9088	0.0000	392	2.1796	2.3985	7.4117	
	NRL(+)	5.5807	0.8656	0.0000	393	2.1010	2.4271	6.7565	
									0.4565
GNAT2									
	CRX(+)	5.2335	1.0998	0.0000	549	2.0192	1.8359	4.3965	
	NRL(+)	5.2272	1.0315	0.0000	532	1.9871	1.9265	4.9988	
									0.2226

Table 1: A snapshot of highly expressed genes in both T+N+R+ (NRL(+)) and T+C+R+ (CRX(+)) cells on a logarithmic scale. Expression values were evaluated and compared.

Nine genes are only in the top 100 genes of the T+C+R+ group (**table 2)**. Genes that are not differentially expressed in the T+N+R+ group are *Syt4*, *Prkca* and *Lmo4*. *Syt4* gene creates SYT-4, a calcium sensor acting in the horizontal cells of the outer plexiform layer[16], *Prkca* is typically found in bipolar cells and is involved in signal transduction and termination[17], and *Lmo4* is a transcription cofactor the helps with the development of amacrine cells in the retina[18]. None of these genes are indicative of rod photoreceptors; however, all these genes are typically expressed in the retina. This means that these genes are not likely to be novel to the gene regulatory network of photoreceptor differentiation.

The other genes that table 2 indicate may be of some interest include *Lhx4* and *Fam19a3*. The p-values for these genes are below 0.05 indicating that the difference in gene expression between T+N+R+ and T+C+R+ is significant. *Lhx4* is a LIM-homeodomain transcription factor that is expressed in bipolar cells[19]. Because this gene is known to have a function in the retina, it is not of interest to us. *Fam19a3* is shown in table 2 to have a greater expression in the T+C+R+ group than the T+N+R+ group. This means that the gene is more likely to be working on the *Crx* pathway. The gene expression is only 1.1x higher in T+C+R+, but since the groups share cells a 1.1x higher expression is significant. For this reason, we looked at the isolated T+N-C+R+ group in section 3.1.2 to determine how much *Crx* influences *Fam19a3* expression. *Fam19a3* does not have a known function, however it has been linked to *Nr2e3* expression in retinas lacking *Nrl* expression in mice[20]. *Nr2e3* is a rod-specific transcriptional regulator deriving from *Nrl* and Cheng et al found that *Nr2e3* expression without deriving from *Nrl* suppresses cone differentiation[20]. *Fam19a3* is not mentioned to have a significant

impact on the system, however our finding could indicate that *Fam19a3* is important in the *Top2b* → *Crx* → *Rho* pathway and the link to the Cheng paper indicates that *Nrl* expression is not necessary for *Fam19a3* expression.

Table 2: Highly expressed Genes in TOP2B(+)CRX(+)RHO(+) Only on a logarithmic scale

		Maximum	Mean	Minimum	Expressed in_Cells	Variance	Dispersion Measure	zScore	p value
SYT4									
	CRX(+)	4.2462	0.8882	0.0000	572	1.2025	1.3539	1.8503	
	NRL(+)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
									N/A
PRKCA									
	CRX(+)	5.5160	0.7644	0.0000	560	1.0667	1.3955	2.0717	
	NRL(+)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
									N/A
LHX4									
	CRX(+)	4.2618	0.6599	0.0000	404	1.1434	1.7325	3.8662	
	NRL(+)	4.0519	0.5675	0.0000	370	1.0072	1.7747	3.6406	
									0.0245
LMO4									
	CRX(+)	4.2524	0.6546	0.0000	482	0.8969	1.3701	1.9366	
	NRL(+)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
									N/A
TRNP1									
	CRX(+)	4.1031	0.6117	0.0000	436	0.9026	1.4755	2.4978	
	NRL(+)	3.9680	0.5631	0.0000	434	0.8132	1.4441	2.0617	
									0.1848
TMEM215									
	CRX(+)	3.5513	0.5947	0.0000	407	0.8845	1.4873	2.5604	
	NRL(+)	3.5454	0.5314	0.0000	386	0.8079	1.5203	2.4252	
									0.0824
FAM19A3									
	CRX(+)	4.4970	0.5892	0.0000	445	0.8217	1.3946	2.0669	
	NRL(+)	4.3034	0.5173	0.0000	422	0.7201	1.3919	1.8121	
									0.0398
NNAT									
	CRX(+)	5.8775	0.5582	0.0000	342	1.1202	2.0070	5.3274	
	NRL(+)	5.7855	0.5266	0.0000	349	1.0264	1.9490	4.4730	
									0.4421
RUNX1T1									
	CRX(+)	4.6202	0.5564	0.0000	428	0.7601	1.3660	1.9145	
	NRL(+)	4.6140	0.5476	0.0000	441	0.7531	1.3753	1.7329	
									0.7978

Table 2: The nine genes that are in the top 100 highest expressed genes for the TOP2B(+)CRX(+)RHO(+) group but not the TOP2B(+)NRL(+)RHO(+) group. N/A rows indicate that the gene is not differentially expressed in the TOP2B(+)NRL(+)RHO(+) group.

The nine genes that are exclusively expressed in the top 100 genes of the T+N+R+ group are shown in **table 3**. *Syt11* and *Jun* are not differentially expressed in

the T+C+R+ group. *Sy11* is a known to be expressed in ganglion cells[21] and *Jun* a proto-oncogene that has been detected in all photoreceptors[22]. These genes are expected to be expressed in the retina so further analysis is not necessary. Genes that have a significant difference in gene expression between T+N+R+ and T+C+R+ groups include *Sparcl1* and *Stmn2*. *Sparcl1* is expressed in the retinal ganglion cell layer[23] and *Stmn2* is also expressed in retinal ganglion cells[23]. These genes do not give any new insight to the gene regulatory pathway for photoreceptors.

Table 3: Highly Expressed Genes in TOP2B(+)NRL(+)RHO(+) Only on a logarithmic scale

		Maximum	Mean	Minimum	Expressed in_Cells	Variance	Dispersion Measure	zScore	p value
SYT11									
	CRX(+)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	NRL(+)	4.3766	1.0052	0.0000	684	1.2329	1.2265	2.2601	
									N/A
SPARCL1									
	CRX(+)	5.6649	0.4770	0.0000	280	1.0189	2.1362	6.0153	
	NRL(+)	5.3578	0.5899	0.0000	352	1.2313	2.0874	5.1341	
									0.0071
SPOCK3									
	CRX(+)	4.6666	0.5508	0.0000	354	0.9707	1.7622	4.0239	
	NRL(+)	4.4222	0.6205	0.0000	408	1.0764	1.7348	3.4498	
JUN									0.0819
	CRX(+)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	NRL(+)	6.3035	0.7307	0.0000	593	1.0001	1.3687	1.7011	
									N/A
RPH3A									
	CRX(+)	4.4962	0.5059	0.0000	329	0.9052	1.7893	4.1684	
	NRL(+)	5.0442	0.5789	0.0000	376	1.0402	1.7968	3.7458	
									0.0613
CACNG4									
	CRX(+)	4.1665	0.5411	0.0000	361	0.9102	1.6822	3.5981	
	NRL(+)	4.1604	0.5994	0.0000	415	0.9712	1.6204	2.9036	
									0.1289
DNER									
	CRX(+)	4.2876	0.5309	0.0000	362	0.8641	1.6276	3.3075	
	NRL(+)	4.2814	0.5867	0.0000	408	0.9498	1.6191	2.8972	
									0.1390
STMN2									
	CRX(+)	5.0765	0.4543	0.0000	271	1.0226	2.2510	6.6263	
	NRL(+)	5.4127	0.5821	0.0000	335	1.3587	2.3342	6.3126	
									0.0031
NSG2									
	CRX(+)	3.7875	0.5188	0.0000	395	0.7218	1.3911	2.0486	
	NRL(+)	3.8586	0.5752	0.0000	442	0.8041	1.3978	1.8402	
									0.103

Table 3: The nine genes that are in the top 100 highest expressed genes for the T+N +R + group but not the T+C+R+ group. N/A rows indicate that the gene is not differentially expressed in the T+C+R+ group.

3.1.2 Significant differences in gene expression between T+C-N+R+ and T+C+N-R+ cells

To clear up which gene is having a greater impact on rod differentiation, we isolated the cells that have no *Crx* expression in the T+N+R+ group, resulting in 175 cells, and the cells that have no *Nrl* expression in the T+C+R+ group, resulting in 118 cells. This allowed for us to see the impact of *Crx* and *Nrl* on *Rho* without having the other gene contributing to the expression. In these cells, we see a significant difference in *Rho* expression, with the *Nrl* group having a 1.5x higher log expression **(Table 4, Figure 5)**. This suggests that *Nrl* has a stronger relationship with rod differentiation. A comparison of the top 100 genes expressed in each of these group also supports this claim. The T+N+R+ group contains within its top 100 genes, genes that are known to be markers for rod photoreceptors, RCVRN and PDE6G. It should be noted that the cells in the isolated groups do not have as strong of a relationship with the marker genes of rods as the cells that have expression of both *Nrl* and *Crx*. For this reason, the conclusion that both *Nrl* and *Crx* are needed for proper rod differentiation and maturation.

Very few cells expressed *Opn1sw* and *Opn1mw* genes in either T+N+C-R+ or T+N-C+R+, with the expression of both genes being similar in the isolated groups. *Gnat2* has a higher expression in the T+N-C+R+ group. This suggests that the *Crx* gene has a stronger relationship with *Gnat2* than *Nrl*. This means that *Crx* is also involved to some degree in cone differentiation. Since we are primarily interested in rod differentiation and

a majority of cells in T+N+R+/T+C+R+ groups are the same, we will use T+N+R+ clustering results for analysis in section 3.1.3. Ultimately, both NRL and CRX are needed for the best differentiation of rod photoreceptors which is supported by literature [10].

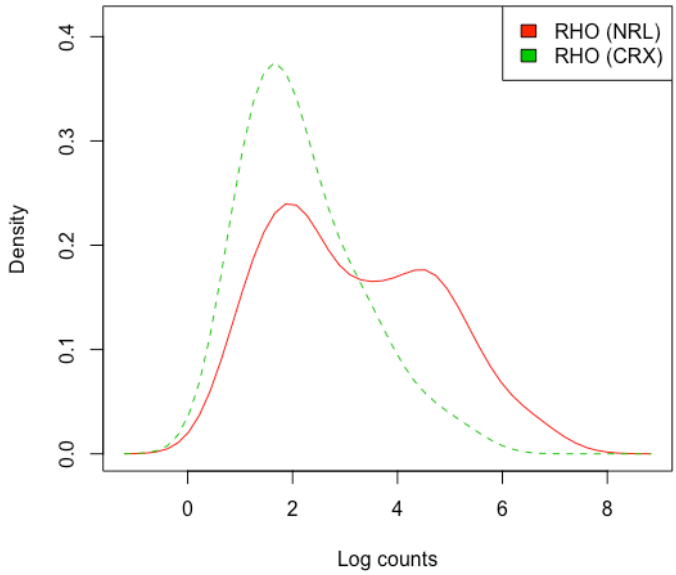

Figure 5: A kernel density plot showing the expression of *Rho* in the isolated groups from the T+N+C-R+ and T+N-C+R+ groups. The x-axis is the log count expression and the y-axis is the "density" or relative likelihood a cell will be found in that space. The red line indicates the expression of RHO in the NRL+ isolated group and the green line indicates the expression of RHO in the CRX+ isolated group.

Table 4: Significant Genes in isolated all positive groups

		Maximum	Mean	Minimum	Expressed in _ Cells	p value
RHO						
	CRX(+)	5.4156	2.1833	0.5582	118	
	NRL(+)	7.1607	3.1821	0.4639	175	
						6.66E-10
TOP2B						
	CRX(+)	3.7080	1.9423	0.6605	118	
	NRL(+)	3.8090	1.8844	0.4499	175	
						0.4772
NRL						
	CRX(+)	0.0000	0.0000	0.0000	0	
	NRL(+)	4.8302	1.6965	0.2943	175	
						N/A
CRX						
	CRX(+)	3.5077	1.5466	0.2960	118	
	NRL(+)	0.0000	0.0000	0.0000	0	
						N/A
SAG						
	CRX(+)	5.5697	1.8015	0.0000	94	
	NRL(+)	6.7705	2.7224	0.0000	153	
						1.23E-06
PDE6G						
	CRX(+)	4.2400	1.0963	0.0000	72	
	NRL(+)	5.9624	1.8072	0.0000	132	
						4.78E-06
RCVRN						
	CRX(+)	4.1214	0.7578	0.0000	55	
	NRL(+)	5.5301	1.7296	0.0000	133	
						5.61E-11
OPN1SW						
	CRX(+)	3.4782	0.1155	0.0000	9	
	NRL(+)	2.9411	0.0586	0.0000	8	
						0.2547
OPN1MW						
	CRX(+)	3.6927	0.1017	0.0000	6	
	NRL(+)	2.1245	0.0544	0.0000	7	
						0.3652
GNAT2						
	CRX(+)	3.6927	0.3438	0.0000	25	
	NRL(+)	3.4620	0.0951	0.0000	8	
						0.001085

Table 4: Significant genes in the T+N+C-R+ and T+N-C+R+ cells.

Of the top 100 genes, only 67 are shared between the isolated groups. This means that there are 66 genes that can potentially provide insight into either the *Top2b* → *Crx* → *Rho* or *Top2b* → *Nrl* → *Rho* pathways, with 33 genes expressed exclusively in the top 100 genes of either group. Only genes that do not have an identified role in the retina were looked into for further expression analysis. 14 genes remained, see **tables 5 and 6** for log normalized gene expression. In **table 5**, *Fam19a3* is highly expressed in both the isolated group of T+N-C+R+ and in the whole T+C+R+ group. This gene is expressed in 41 cells in the T+N-C+R+ group but only 18 cells in the T+N+C-R+ group. The difference in expression is found to be significant by the student's t-test. Average expression is almost 5x higher in the T+N-C+R+ group for *Fam19a3*. This gene is therefore a prime candidate for interacting in the *Top2b* → *Crx* →*Rho* pathway. BC030499 also has a significant difference between the isolated groups that is found to be statistically significant. The expression of this gene is slightly higher in the T+N-C+R+ group. However, this gene expression was not strong enough to have an impact on the T+C+R+ group as a whole. This gene is not as promising as *Fam19a3* but will be looked into further in section 3.2.

Table 6 has gene expression of genes that are only expressed in the top 100 genes of T+N+C-R+ group. These genes do not have any known function in the retina and therefore are potential regulatory genes for the *Top2b* → *Nrl* → *Rho* network. Of the eight genes shown, six of them prove to have a significant difference in expression between isolated groups. All six of these genes have a higher expression in the T+N+C-R+ group, but the genes I would like to highlight are *Gria3*, *Rph3a* and *Cdk14*. *Gria3* has a two-fold higher average log count expression in T+N+C-R+ cells, as well as is

expressed in almost three times as many cells. A similar result is found for *Cdk14* which also has 2x higher expression and three times as many cells expression the gene. *Rph3a* does not have quite as dramatic a difference in gene expression, but was among the highest expressed gene in T+N+R+. This means that the difference in expression of *Rph3a* between T+N+R+ and T+C+R+ groups was strong enough for this gene to be noted.

Table 5: Significant Genes Only Expressed in the Isolated CRX(+) Group

		Maximum	Mean	Minimum	Expressed in_Cells	p value
FAM19A3						
	CRX(+)	4.4970	0.7218	0.0000	41	
	NRL(+)	2.8261	0.1515	0.0000	18	
						6.47E-07
NNAT						
	CRX(+)	5.8775	1.1818	0.0000	54	
	NRL(+)	5.5274	0.7506	0.0000	61	
						0.01
B3GALT2						
	CRX(+)	4.3682	1.2748	0.0000	69	
	NRL(+)	3.7626	0.5637	0.0000	53	
						4.58E-07
GPR179						
	CRX(+)	4.1848	1.2424	0.0000	62	
	NRL(+)	4.2449	0.4433	0.0000	35	
						1.15E-07
BC030499						
	CRX(+)	3.8779	1.1323	0.0000	70	
	NRL(+)	3.8608	0.8220	0.0000	79	
						0.01976
TMEM215						
	CRX(+)	3.4487	0.9523	0.0000	54	
	NRL(+)	3.4829	0.3723	0.0000	33	
						4.82E-06

Table 5: Significant genes that are only expressed in the top 100 genes of TOP2B(+)NRL(-)CRX(+)RHO(+) cells.

Table 6: Significant Genes Only Expressed in the Isolated NRL(+) Group

		Maximum	Mean	Minimum	Expressed in _ Cells	p value
C1QL1						
	CRX(+)	4.4536	0.8622	0.0000	37	
	NRL(+)	4.5832	1.2811	0.0000	78	
						0.0175
RPH3A						
	CRX(+)	3.4304	0.8347	0.0000	50	
	NRL(+)	5.0442	1.2797	0.0000	97	
						0.001739
TPM3						
	CRX(+)	N/A	N/A	N/A	N/A	
	NRL(+)	6.4739	1.2038	0.0000	110	
						N/A
CDK14						
	CRX(+)	3.7453	0.6246	0.0000	39	
	NRL(+)	4.4616	1.1258	0.0000	103	
						0.000164
GRIA3						
	CRX(+)	3.6270	0.4780	0.0000	30	
	NRL(+)	4.4447	0.9972	0.0000	89	
						2.87E-05
FRMD5						
	CRX(+)	4.0301	0.7979	0.0000	51	
	NRL(+)	3.6498	0.9842	0.0000	89	
						0.1589
AI593442						
	CRX(+)	4.6516	0.5635	0.0000	33	
	NRL(+)	3.9364	0.8436	0.0000	69	
						0.03532
6330403K07RIK						
	CRX(+)	3.4172	0.5758	0.0000	41	
	NRL(+)	3.5523	0.8333	0.0000	85	
						0.02067

Table 6: Gene expression of genes that are expressed only in the TOP2B(+)NRL(+)CRX(-)RHO(+) top 100 genes.

3.1.3 Clustering of T+N+R+ cells supports claim that these cells are photoreceptor cells

Since the difference in number of cells between the T+N+R+ and T+C+R+ groups is so small, we can utilize the *Nrl* gene regulatory network for the remainder of our analysis because *Nrl* has a higher chance of producing rod cells. Both groups produced 13 clusters identified by SC3. A heatmap was produced explaining the relationship between all cells in the groups (**Figure 6)**. All clusters were not clearly defined, with the exception of cluster 11. This cluster produced 76 marker genes (see appendix), but none were marker genes for rods. Other clusters that are clearer are 1, 2 and 3. Cluster 1's marker genes include *Prdm1*, *Arr*3 and *Opn1mw*, indicating that these cells are cone photoreceptor cells. *Prdm1* is a gene that prevents a photoreceptor cell from specifying into a bipolar cell[13]. *Arr3* is a known retinal cone arrestine. Cluster 2 has some similarities with cluster 1 in that it appears to be cone photoreceptors. Cluster 2's significant marker genes include *Opn1sw*, *Gnat2*, and *Pde6h* (see appendix for more). *Pde6h* is a cone-specific phosphodiesterase[15]. Cluster 3 is the only cluster that is very clearly rod photoreceptors, with *Pde6a* and *Crx* as marker genes. The remaining clusters are also rod photoreceptors but gene expression is too similar to separate the cells. This validates SC3's clustering algorithm from its ability to tease out cone photoreceptor cells from the rod photoreceptor cell. This means that later in section 3.2 the results found there are valid.

Genes that were found to be significant in the T+N-C+R+ group, *Fam19a3* and *Bc030499*, are marker genes for cluster 6 and 7 for T+N+R+. This group is not clearly defined, likely because these cells are photoreceptor cells. This proves that a stronger algorithm is needed to tease out potential subgroups in rod photoreceptors. This also

supports the hypothesis that *Fam19a3* and *Bc030499* are involved in the differentiation of rod photoreceptors.

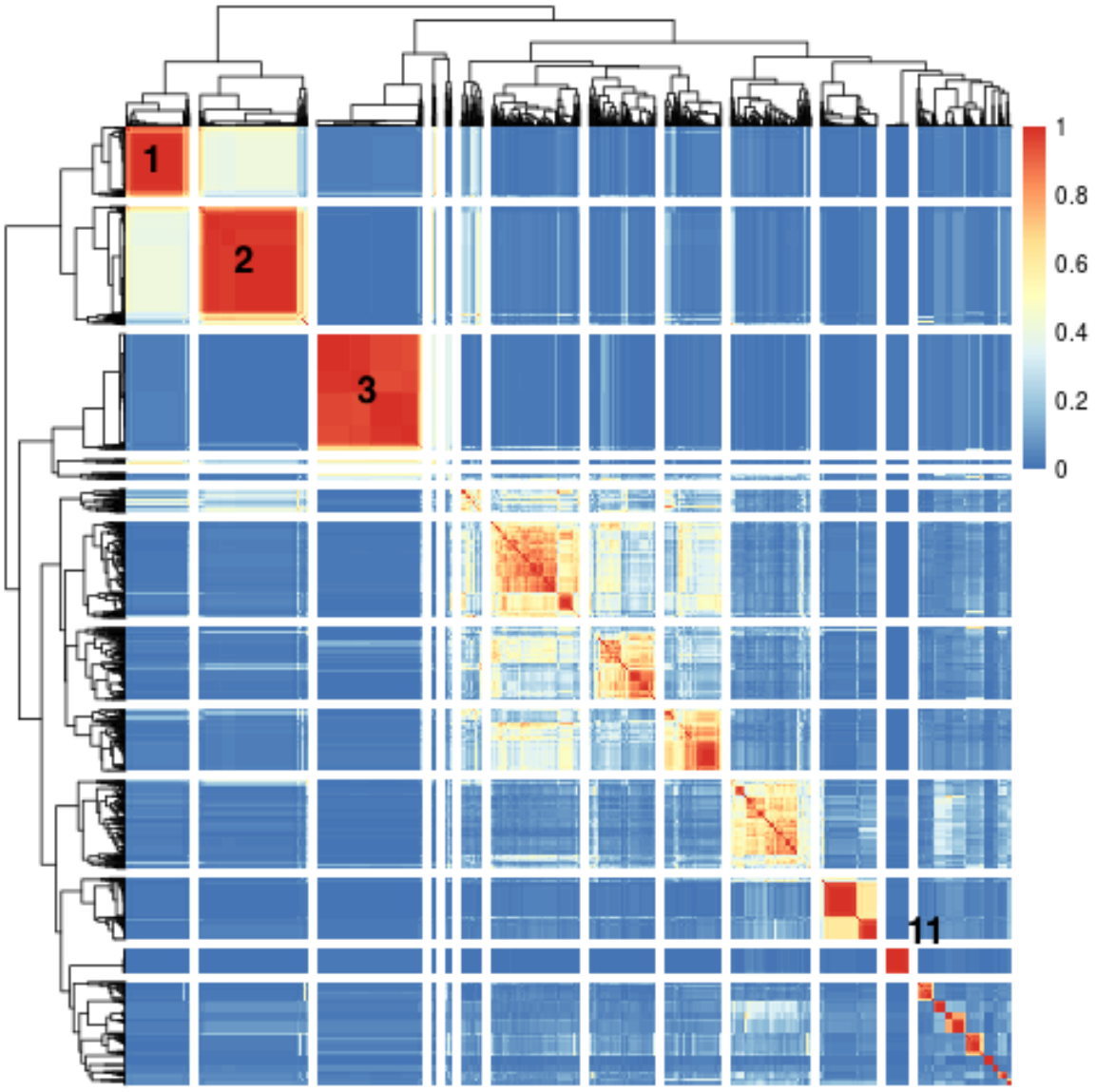

Figure 6: SC3 heatmap plot of cells in TOP2B(+)NRL(+)RHO(+).The x and y axes are cells, ideally there will be a clear red diagonal indicating similarity between nearby cells. White space in the plot indicate different clusters. Darker red indicates a stronger relationship. Darker blue means that there is no relationship between cells

3.2 Comparison of T+N+R+/T+C+R+ and T-N+R+/T-C+R+ cells shows Top2b expression promotes expression of rod marker genes

Top2b is expected to be expressed in all post-mitotic cells[7]. However, there exists several hundred cells in this dataset that do not have any expression of *Top2b* in this dataset. Here, we analyze the differences in gene expression and clusters between the *Top2b* positive and *top2b* negative photoreceptor cells.

3.2.1 Comparison of T+N+R+ and T-N+R+ cells shows stronger gene expression of rod marker genes

The top highest expressed genes in both T+N+R+ and T-N+R+ include rod photoreceptor marker genes *Rho*, *Sag, Pde6g* and *Rcvrn* (**figure 7 and 8)**. This indicates that both groups contain a large amount photoreceptor cells. No significant difference can be discerned from the figures alone, so we looked at the expressions of these genes to determine if there is any difference in how these genes are expressed in the *Top2b* positive and negative groups (**table 7)**. Despite the groups having nearly identical top 10 genes, it can be seen that *Rho, Nrl, Crx, Sag, Pde6g* and *Rcvrn* expression is higher in cells that express *Top2b*. This is confirmed to be statistically significant by the students t-test, shown by the p values. In T+N+R+ cells, the percentage of cells that have the rod marker genes expressed is always higher than cells without *Top2b* expression. It can be inferred that *Top2b*, while not necessary for differentiation into rods, promotes the quality and number of cells that differentiate into rods. However, no difference was found in cone marker genes, likely because NRL is involved more directly with rod differentiation than in cone differentiation.

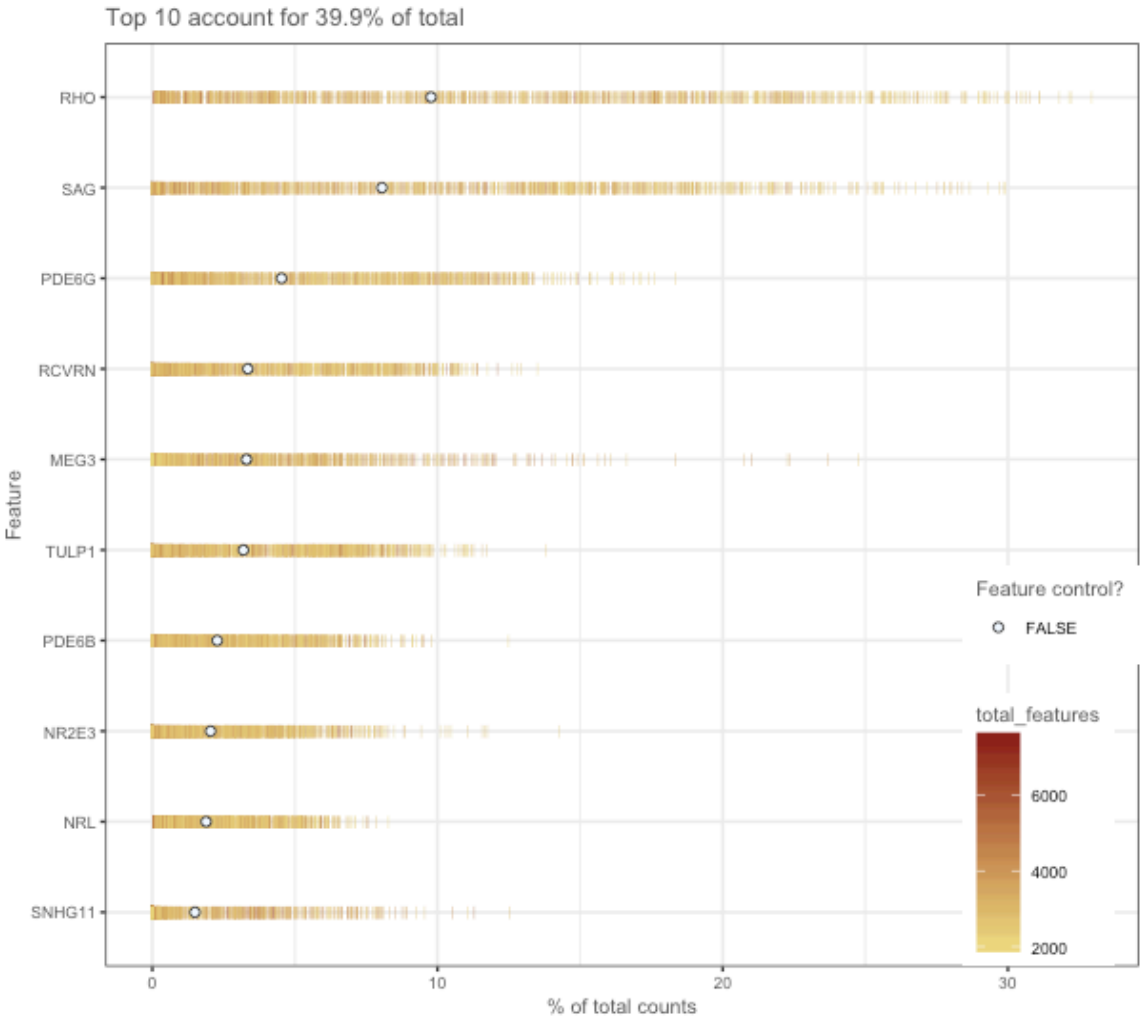

Figure 7: A plot of the top 10 highest expressed genes in TOP2B(+)NRL(+)RHO(+) group. The y axis is the name of the gene and the x axis indicates the percentage of total counts in each cell the gene accounts for. Each line indicates a cell. The color of the line indicates the total genes that are expressed in that cell. Since we are only looking at differentially expressed genes, the total features are always below 2000.

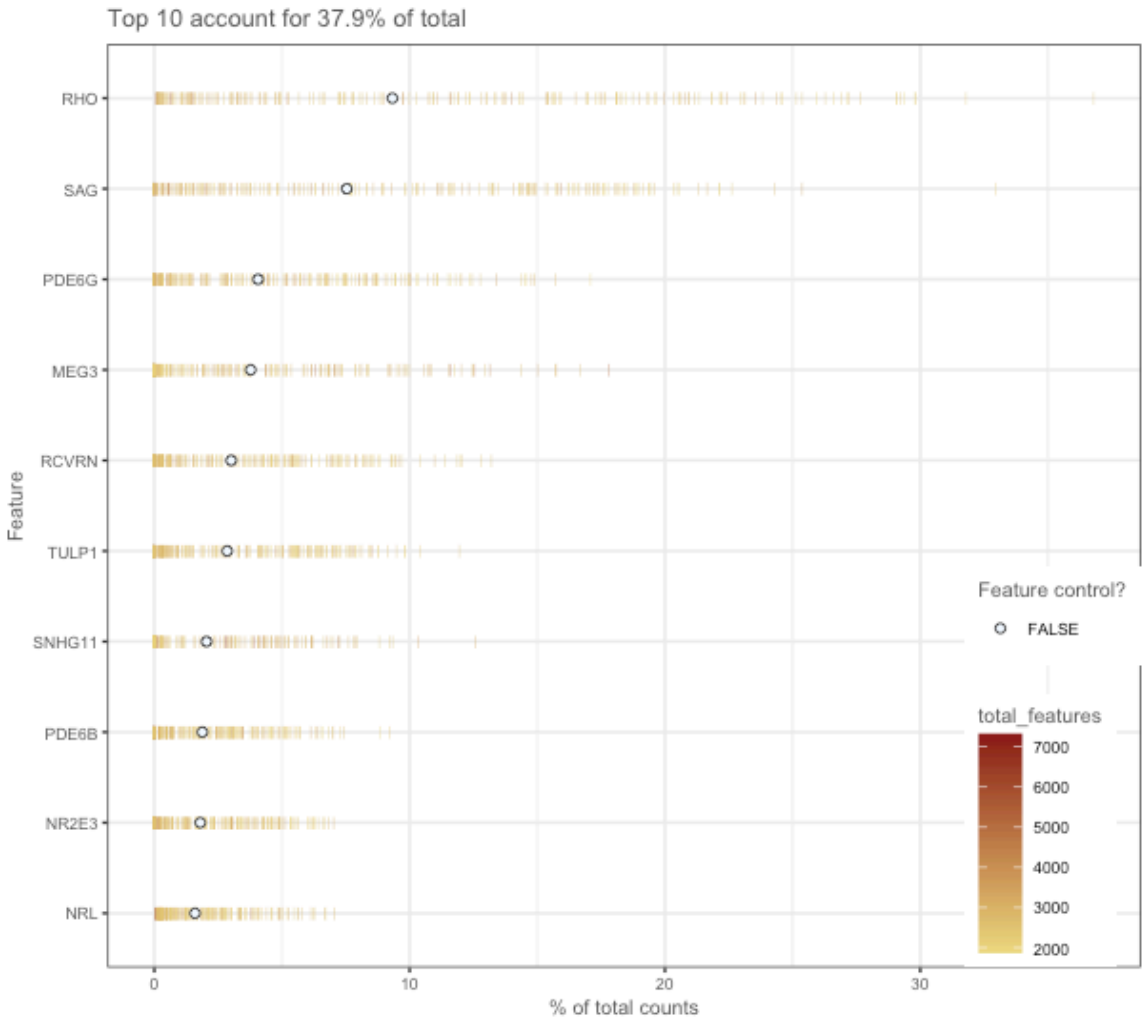

Figure 8: A plot of the top 10 highest expressed genes in TOP2B(-)NRL(+)RHO(+) group. The y axis is the name of the gene and the x axis indicates the percentage of total counts in each cell the gene accounts for. Each line indicates a cell. The color of the line indicates the total genes that are expressed in that cell. Since we are only looking at differentially expressed genes, the total features are always below 2000.

Table 7: Photoreceptor Marker Genes Expression in TOP2B(+/-)NRL(+)RHO(+) Groups

		Maximum	Mean	Minimum	Expressed in _Cells	Expressed in % of Total Cells	Variance	Dispersion Measure	zScore	p value
RHO										
	TOP2B(+)	7.7533	5.3973	0.4639	1306	100.00%	2.6921	0.4988	-0.5870	
	TOP2B(-)	7.5900	4.8236	0.5161	212	100.00%	3.4223	0.7095	0.1807	
										2.93E-05
TOP2B										
	TOP2B(+)	4.1932	2.0297	0.4499	1306	100.00%	0.4335	0.2136	-1.7027	
	TOP2B(-)	#N/A	#N/A	#N/A	#N/A	0.00%	#N/A	#N/A	#N/A	
										N/A
NRL										
	TOP2B(+)	5.7144	3.2130	0.2943	1306	100.00%	1.6876	0.5252	-0.4834	
	TOP2B(-)	5.0357	2.6760	0.2023	212	100.00%	1.6323	0.6100	-0.1777	
										3.61E-08
CRX										
	TOP2B(+)	4.5730	2.3431	0.0000	1131	86.60%	1.5632	0.6672	0.0719	
	TOP2B(-)	4.4136	1.8456	0.0000	154	72.64%	1.9016	1.0303	1.3364	
										1.41E-06
SAG										
	TOP2B(+)	7.5650	5.1152	0.0000	1275	97.63%	3.0033	0.5871	-0.2413	
	TOP2B(-)	6.9685	4.4994	0.0000	200	94.34%	3.7010	0.8225	0.5880	
										1.69E-05
PDE6G										
	TOP2B(+)	7.2312	4.1431	0.0000	1241	95.02%	3.1658	0.7641	0.4512	
	TOP2B(-)	6.7545	3.5736	0.0000	192	90.57%	3.6099	1.0101	1.2637	
										5.84E-05
RCVRN										
	TOP2B(+)	6.4847	3.7816	0.0000	1226	93.87%	2.8626	0.7570	0.4232	
	TOP2B(-)	6.2456	3.1503	0.0000	185	87.26%	3.5130	1.1151	1.6419	
										6.24E-06
OPN1SW										
	TOP2B(+)	7.0337	0.9980	0.0000	354	27.11%	3.3897	3.3965	11.3866	
	TOP2B(-)	5.6323	0.9483	0.0000	56	26.42%	2.9229	3.0824	9.3539	
										0.6979
OPN1MW										
	TOP2B(+)	5.5807	0.8656	0.0000	393	30.09%	2.1010	2.4271	6.7565	
	TOP2B(-)	4.4940	0.7668	0.0000	60	28.30%	1.8089	2.3590	6.1153	
										0.3273
GNAT2										
	TOP2B(+)	5.2272	1.0315	0.0000	532	40.74%	1.9871	1.9265	4.9988	
	TOP2B(-)	4.6183	0.8630	0.0000	77	36.32%	1.6237	1.8816	3.9782	
										0.07964

Table 7: Photoreceptor marker genes expression in TOP2B(+/-)NRL(+)RHO(+) groups.

In section 3.1.2, we found 14 genes that have no known function in the retina but were highly expressed in the isolated all positive groups. To see if these genes have any relation to the *Top2b*, we looked at the differences in expression shown in **table 8**. Of the top 100 highest expressed genes in both groups, *B3galt2, Gpr179, Bc030499* and *Frmd5* were found to be only expressed in the T+N+R+ group. Of those genes, *B3galt2, Gpr179* and *Bc030499* have statistically significant differences between groups. These genes

Table 8: Significant Genes found from Section 3.1.2 Expressed in TOP2B(+/-)NRL(+)RHO(+)

		Maximum	Mean	Minimum	Expressed in _Cells	Expressed in % of Total Cells	Variance	Dispersion Measure	zScore	p value
FAM19A3										
	TOP2B(+)	4.3034	0.5173	0.0000	422	32.31%	0.7201	1.3919	1.8121	
	TOP2B(-)	3.4952	0.3079	0.0000	44	20.75%	0.4358	1.4152	1.8905	
										5.16E-05
NNAT										
	TOP2B(+)	5.7855	0.5266	0.0000	349	26.72%	1.0264	1.9490	4.4730	
	TOP2B(-)	3.3296	0.4035	0.0000	51	24.06%	0.6327	1.5679	2.5742	
										0.04578
B3GALT2										
	TOP2B(+)	4.3057	0.5857	0.0000	430	32.92%	0.8668	1.4799	2.2326	
	TOP2B(-)	3.6077	0.4511	0.0000	54	25.47%	0.7355	1.6304	2.8536	
										0.03714
GPR179										
	TOP2B(+)	4.7306	0.5768	0.0000	358	27.41%	1.0898	1.8893	4.1879	
	TOP2B(-)	3.9159	0.4187	0.0000	47	22.17%	0.7414	1.7705	3.4812	
										0.01686
BC030499										
	TOP2B(+)	4.6130	0.6554	0.0000	479	36.68%	0.9557	1.4581	2.1284	
	TOP2B(-)	3.6106	0.4625	0.0000	64	30.19%	0.6230	1.3471	1.5855	
										0.001585
TMEM215										
	TOP2B(+)	3.5454	0.5314	0.0000	386	29.56%	0.8079	1.5203	2.4252	
	TOP2B(-)	3.1190	0.3641	0.0000	48	22.64%	0.5286	1.4518	2.0541	
										0.002911
C1QL1										
	TOP2B(+)	4.5832	0.4272	0.0000	226	17.30%	1.0524	2.4638	6.9320	
	TOP2B(-)	4.2122	0.4876	0.0000	42	19.81%	1.1422	2.3426	6.0423	
										0.4433
RPH3A										
	TOP2B(+)	5.0442	0.5789	0.0000	376	28.79%	1.0402	1.7968	3.7458	
	TOP2B(-)	3.4291	0.5495	0.0000	73	34.43%	0.7161	1.3031	1.3885	
										0.6491
TPM3										
	TOP2B(+)	6.4739	0.4993	0.0000	451	34.53%	0.6915	1.3847	1.7779	
	TOP2B(-)	5.0299	0.4171	0.0000	66	31.13%	0.5764	1.3819	1.7412	
										0.15
CDK14										
	TOP2B(+)	4.4616	0.3996	0.0000	316	24.20%	0.6523	1.6321	2.9595	
	TOP2B(-)	3.3991	0.4261	0.0000	58	27.36%	0.6204	1.4558	2.0722	
										0.6513
GRIA3										
	TOP2B(+)	4.4447	0.3404	0.0000	254	19.45%	0.5984	1.7580	3.5607	
	TOP2B(-)	2.9541	0.4229	0.0000	52	24.53%	0.6483	1.5331	2.4184	
										0.1653
FRMD5										
	TOP2B(+)	3.9030	0.5356	0.0000	424	32.47%	0.7628	1.4242	1.9663	
	TOP2B(-)	3.0428	0.4244	0.0000	59	27.83%	0.5952	1.4025	1.8336	
										0.05702
AI593442										
	TOP2B(+)	3.9520	0.2955	0.0000	186	14.24%	0.6326	2.1411	5.3907	
	TOP2B(-)	4.1516	0.3224	0.0000	31	14.62%	0.7159	2.2207	5.4963	
										0.6653
6330403K07RIK										
	TOP2B(+)	3.6407	0.3506	0.0000	301	23.05%	0.4977	1.4196	1.9445	
	TOP2B(-)	4.2208	0.3858	0.0000	51	24.06%	0.6022	1.5610	2.5429	
										0.5357

Table 8: Expression table of genes that possibly contribute to the TOP2B → NRL -- > RHO regulatory pathway. Italicized and bolded genes are genes that are found exclusively in the top 100 genes of TOP2B(+) cells.

are regularly expressed more highly in T+N+R+ cells. The *Bc030499* gene is expressed in 6.5% more cells in T+N+R+ groups and is on average 1.4x higher in logarithmic expression. *B3galt2* is expressed in 7% more cells and is on average 1.3x higher in logarithmic expression and *Gpr179* is expressed in 5% more cells and 1.4x higher in logarithmic expression. These genes and their expressions in these groups indicate that *Top2b* could be regulating their expression on some level. However, *Bc030499, B3galt2* and *Gpr179* were found to be more highly expressed in T+C+R+ cells, which we will look into next.

3.2.2 Comparison of T+C+R+ and T-C+R+ cells shows stronger gene expression of rod marker genes but no difference in cone marker gene expression

Expression of rod photoreceptor marker genes in T+C+R+ and T-C+R+ groups is strong and once again strongly suggests that these are photoreceptor cells. Differences in gene expression show (**table 9)** that rod specific marker genes (*Rho, Sag, Pde6g, Rcvrn*) are more highly expressed in cells with *Top2b* expression. These results are significant based on the p-values obtained from the students t-test. Once again, these computational results show that *Top2b*, while perhaps not vital for rod differentiation, does enhance and prove maturity of rod cells. The cells that do not have *Top2b* expression could still be in the mitotic phase and have not reached full maturity. However, differences in cone marker gene expression (*Opn1sw, Opn1mw* and *Gnat2*) are not affected by *Top2b* expression.

Genes of interest that were found in section 3.1.2 from isolated groups are once again looked at. *Fam19a3* and *Nnat* were found to be in the top 100 highly expressed genes in T+C+R+ group but not in the T-C+R+ group. The average expression of

Fam19a3 in the T+C+R+ group is 1.5x higher than in the T-C+R+ group, as well as having 10% more cells expression. The bioinformatics analysis suggests that *Fam19a3* is a prime candidate for participating in the *Top2b* → *Crx* →*Rho* gene regulatory system. Another candidate is *Nnat*. *Nnat* was not found to have a significant role in either *Nrl* or *Crx* specific pathways, seen in section 3.1.2, however in **tables 8 and 10** differences in expression between *Top2b*(+) and *Top2b*(-) cells are seen. There is a more significant expression of *Nnat* with the T+C+R+ cells, having a p value of 0.003995 compared to the T+N+R+ group's p value of 0.045. The computational results suggests that *Nnat* has some type of relationship with *Top2b* and *Crx*. The average expression of *Nnat* in T+C+R+ cells is 1.5x higher than that of T-C+R+ cells. T+C+R+ cells also have a 10% higher chance of expressing *B3galt2* with a 1.4x higher expression average. Other significant differences in expression include *Gpr179* and *Bc030499*. All these genes were originally found from the isolated T+N-C+R+ group of cells, and they all have a relationship with *Top2b* expression. These results indicate that these genes work more directly with *Crx* than *Nrl*.

Table 9: Photoreceptor Marker Genes in TOP2B(+/-)CRX(+)RHO(+)

		Maximum	Mean	Minimum	Expressed in _ Cells	% Cells Expressed	p value
RHO							
	TOP2B(+)	7.7597	5.4097	0.5582	1249	100.00%	
	TOP2B(-)	7.5574	5.0173	0.7956	183	100.00%	
							< 2.2e-16
NRL							
	TOP2B(+)	5.7207	3.1270	0.0000	1131	90.55%	
	TOP2B(-)	5.0040	2.6123	0.0000	154	84.15%	
							3.51E-05
CRX							
	TOP2B(+)	4.5792	2.6009	0.2778	1249	100.00%	
	TOP2B(-)	4.3825	2.3387	0.1981	183	100.00%	
							5.69E-04
SAG							
	TOP2B(+)	7.5714	5.1430	0.0000	1216	97.36%	
	TOP2B(-)	6.9360	4.7846	0.0000	180	98.36%	
							9.42E-03
PDE6G							
	TOP2B(+)	7.2377	4.1879	0.0000	1181	94.56%	
	TOP2B(-)	6.7221	3.7857	0.0000	167	91.26%	
							0.006353
RCVRN							
	TOP2B(+)	6.4911	3.7887	0.0000	1148	91.91%	
	TOP2B(-)	6.2134	3.3675	0.0000	163	89.07%	
							0.003555
OPN1SW							
	TOP2B(+)	7.0401	1.0478	0.0000	355	28.42%	
	TOP2B(-)	5.6002	1.0679	0.0000	55	30.05%	
							8.87E-01
OPN1MW							
	TOP2B(+)	5.5870	0.9088	0.0000	392	31.39%	
	TOP2B(-)	4.7061	0.9330	0.0000	61	33.33%	
							8.33E-01
GNAT2							
	TOP2B(+)	5.2335	1.0998	0.0000	549	43.96%	
	TOP2B(-)	4.5869	1.0810	0.0000	83	45.36%	
							0.8611

Table 9: Photoreceptor Marker genes in TOP2B(+/-)CRX(+)RHO(+) groups.

Table 10: Significant Gene Expression in TOP2B(+/-)CRX(+)RHO(+)

		Maximum	Mean	Minimum	Expressed in _ Cells	% Cells Expressed	p value
FAM19A3							
	TOP2B(+)	4.4970	0.5892	0.0000	445	35.63%	
	TOP2B(-)	3.8881	0.3961	0.0000	47	25.68%	
							0.001899
NNAT							
	TOP2B(+)	5.8775	0.5582	0.0000	342	27.38%	
	TOP2B(-)	3.0513	0.3723	0.0000	41	22.40%	
							0.003995
B3GALT2							
	TOP2B(+)	4.3682	0.6552	0.0000	446	35.71%	
	TOP2B(-)	3.5777	0.4727	0.0000	47	25.68%	
							0.01125
GPR179							
	TOP2B(+)	4.7368	0.6596	0.0000	385	30.82%	
	TOP2B(-)	3.3269	0.4993	0.0000	48	26.23%	
							0.03134
BC030499							
	TOP2B(+)	4.6192	0.6785	0.0000	470	37.63%	
	TOP2B(-)	3.5806	0.5298	0.0000	57	31.15%	
							0.03691
C1QL1							
	TOP2B(+)	4.4536	0.3492	0.0000	185	14.81%	
	TOP2B(-)	4.1813	0.4100	0.0000	34	18.58%	
							0.4181
GRIA3							
	TOP2B(+)	3.7558	0.2619	0.0000	195	15.61%	
	TOP2B(-)	2.6957	0.2881	0.0000	34	18.58%	
							0.6124
AI593442							
	TOP2B(+)	4.6516	0.2444	0.0000	150	12.01%	
	TOP2B(-)	4.1899	0.3583	0.0000	41	22.40%	
							0.6832

Table 10: Possible candidates for TOP2B → CRX → RHO gene regulatory network regulation.

3.2.3 Clustering of T-N+R+ cells shows novel gene is among photoreceptor marker genes

SC3 provides a method of unsupervised clustering that we applied to the T-N+R+ group. The result is 6 clusters shown in **figure 9**. Cluster 2 has all the marker genes for rod photoreceptors including *Rcvrn, Tulp1, Pde6h* and *Crx*, strongly suggesting that this cluster contains photoreceptor cells. Additionally, *Fam19a3* is a marker gene for these

cells. This suggests that *Fam19a3* is also involved in photoreceptor differentiation. The remaining clusters' marker genes contain known marker genes for either bipolar or ganglion cells. See appendix for full list of marker genes for all clusters.

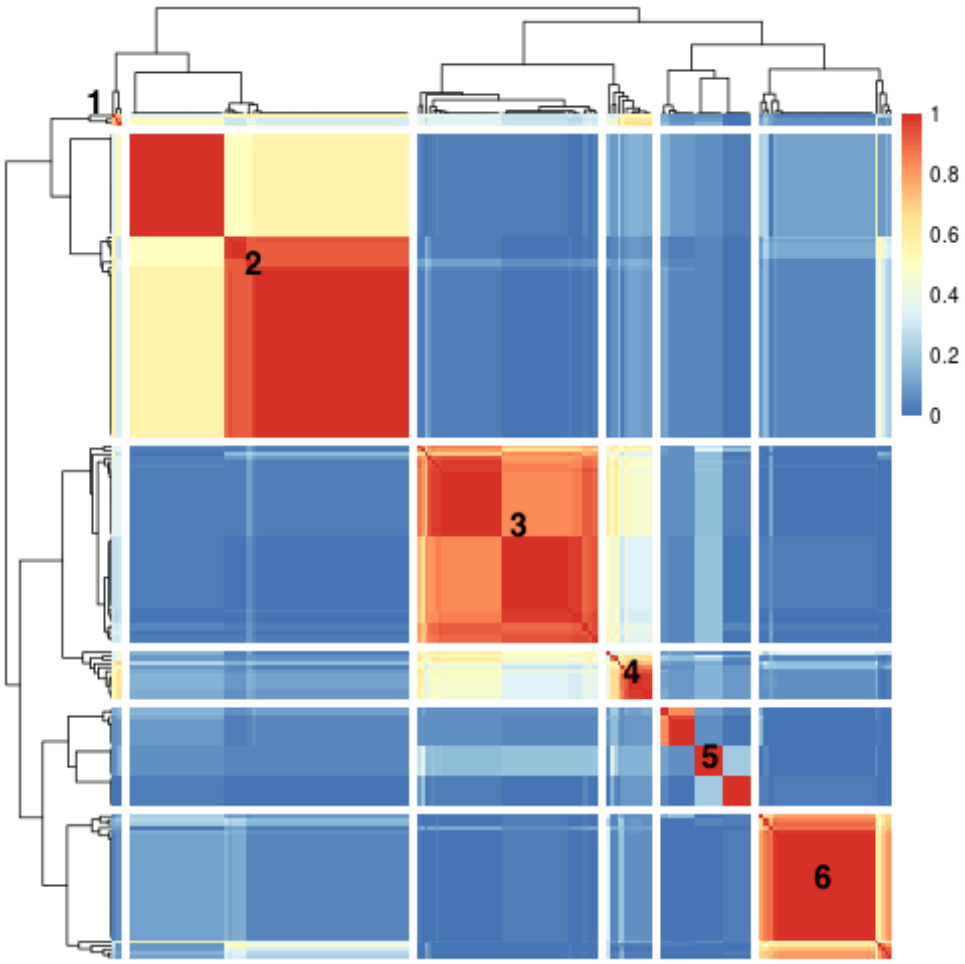

Figure 9: SC3 heatmap plot of cells in TOP2B(-)NRL(+)RHO(+).The x and y axes are cells, ideally there will be a clear red diagonal indicating similarity between nearby cells. White space in the plot indicate different clusters. Darker red indicates a stronger relationship. Darker blue means that there is no relationship between cells

3.2.3 Clustering of T-C+R+ cells shows novel gene is among photoreceptor marker genes

Unsupervised clustering of the T-C+R+ group found only 4 clusters. None of the clusters appear to have clear clustering but cluster 1 contains the marker genes for rod photoreceptor cells, including *Rcvrn, Nrl, Tulp1, Pde6d* and *Nr2e3.* This group contains 80 of the 183 cells in the group and is the largest cluster (**figure 10**). Interestingly, this group also contains *Fam19a3* as a marker gene. This further solidifies our claim that *Fam19a3* could be involved in the *Top2b* → *Crx* → *Rho* gene regulatory pathway. The other remaining clusters contain long lists of marker genes, most of which indicate ganglion or bipolar cell types (see appendix for list).

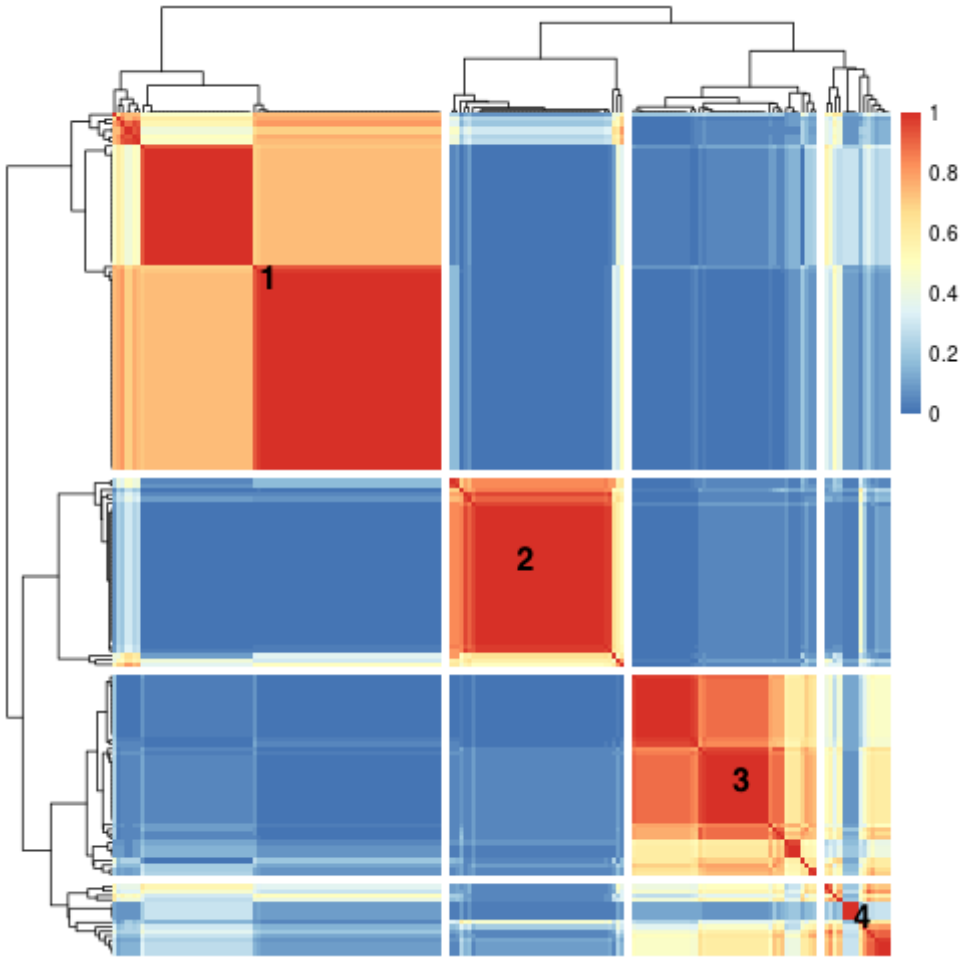

Figure 10: SC3 heatmap plot of cells in TOP2B(-)CRX(+)RHO(+).The x and y axes are cells, ideally there will be a clear red diagonal indicating similarity between nearby cells. White space in the plot indicate different clusters. Darker red indicates a stronger relationship. Darker blue means that there is no relationship between cells

3.3 Conclusions

From the computational results given, we can only guess the role of *Top2b* in rod differentiation. While the results suggest that *Top2b* results in higher *Rho* and other rod photoreceptor marker gene expression, more data is needed to confirm the exact nature of its involvement. These cells are expected to be fully mature and no longer undergoing mitosis at this time point, meaning *Top2b* will be expressed in all cells. However, there are several hundred of cells that do not exhibit any *Top2b* expression. The cells that have the expression profile of a rod photoreceptor but do not have *Top2b* expression have significantly less expression of the photoreceptor marker genes. This means that the cells could still be differentiating at this mature stage. Alternatively, the sequenced data could be of poor quality and these cells may actually have *Top2b* expression, but the transcripts were missing when the cells were sequenced. To confirm the exact contribution of *Top2b* further analysis at several different time points with several datasets would be necessary.

The computational approach of separating the cells based on the gene regulatory pathways proved to provide some novel insights. In a comparison of *Crx* and *Nrl* gene regulatory pathways in photoreceptors, we found that *Fam19a3* is a promising candidate to be involved in the *Top2b* → *Crx* →*Rho* pathway. *Fam19a3* is expressed more highly in T+N-C+R+ cells than in T+N+C-R+ cells. This means that *Fam19a3* is more closely related to *Crx*. Additionally, when comparing *Top2b* negative and positive cells, *Fam19a3* is consistently more highly expressed in *Top2b* positive cells. Other genes that

can potentially be contributing to a photoreceptor gene regulatory pathway includes *Bc030499, Nnat* and *B3galt2*. However, none of these genes have as clear a relationship with photoreceptor expression as *Fam19a3*.

3.4 Future Directions

Here, we used bioinformatics analysis to provide some insight into the photoreceptor gene regulatory network. We found that separating the cells based on the gene regulatory network makes it easier to identify cell types and can provide novel insights. We found that less cells are photoreceptors if *Top2b* is not expressed and that expression of both *Nrl* and *Crx* can upregulate the rod marker genes. We also found a novel gene that could be involved in the *Top2b* → *Crx* → *Rho* pathway. This thesis was focused only on bioinformatics methods on a healthy retina with nothing altered. This means that clearly defining the pathways needs further experiments.

We hope to use similar bioinformatics analysis on our own wet lab experiments of single cell RNA-seq data. We need to use data to analyze the retina at different developmental time points, not just P14 when most cells are already mature. If we can look at the different time points, we will truly be able to see if *Top2b* is contributing to the development of *Rho* and the differentiation of rod photoreceptors. Our lab is currently performing *Top2b* knockout experiments in the retina and working towards following Macosko's Drop-seq method of gathering RNA-seq libraries. Additionally, knockout experiments and staining of *Fam19a3* should be utilized to identify whether or not this gene is significant in the photoreceptor differentiation pathway.

Appendix A

Software Packages

Analysis was performed on the server that works on Ubuntu x86_64-pc-linux-gnu and RStudio. Work was done primarily in R version 3.4.3. R Packages needed include:

- Hexbin_1.27.2
- Sm_2.2-5.4
- Scatterplot3d_0.3-40
- SC3_1.7.7
- Xlsx_0.5.7
- Xlsjars_0.6.1
- rJava_0.9-9
- scater_1.6.2
- ggplot2_2.2.1
- SingleCellExperiment_1.0.0
- SummarizedExperiment_1.8.1
- DelayedArray_0.4.1
- MatrixStats_0.53.0
- Biobase_2.38.0
- GenomicRanges_1.30.1
- GenomeInfoDb_1.14.0
- IRanges_2.12.0
- S4Vectors_0.16.0
- BiocGenerics_0.24.0

Appendix B

Code

B.1 Obtaining Data

From Gene Expression Omnibus, GSE63472 "Drop-Seq analysis of P14 mouse retina single-cell suspension" was downloaded by selecting GSE63472_P14Retina_merged_digital_expression.txt.gz. This was renamed in the server to p14unsorted.txt for simplicity.

B.2 Filtering and Separating Groups

The count matrix was loaded into R and filtered. Genes that had no expression in any of the cells were removed and spike in were labeled. Scater's calculateQCMetrics method was run to simplify the format of the matrix. Cells were filtered out if less than 4000 transcript reads were found. The index location of TOP2B, NRL, CRX and RHO were found. The indices were then used to break the data into 16 different groups based on whether or not each gene was expressed. Code for the NRL branched data is shown below.

```
library(scater)
library(SC3)
#First steps using p14sorted.txt
#p14 <- read.table("P14unsorted.txt", header = TRUE, row.names = 1)
#p14 <- data.matrix(p14)
#p14SCE <- SingleCellExperiment(assays = list(counts = p14))
#Original Data contains 24658 genes and 49300 cells
load("~/Documents/Research Summer 2017/p14sce.RData")

#remove genes that have no expression results in 24071 genes kept
keep_gene <- rowSums(counts(p14SCE)>0)>0
p14SCE <- p14SCE[keep_gene,]

#Labels spike ins
isSpike(p14SCE, "ERCC") <- grepl("^ERCC", rownames(p14SCE))
isSpike(p14SCE,"MT") <- grepl("^MT-",rownames(p14SCE))
```

```
#normalize data by log, results in assay data called "logcounts"
exprs(p14SCE) <- log2(calculateCPM(p14SCE, use.size.factors = FALSE)+1)

#calculateQC metrics and eliminate cells with poor quality
p14SCE <- calculateQCMetrics(p14SCE, feature_controls = list(ERCC =
isSpike(p14SCE,"ERCC"), MT = isSpike(p14SCE,"MT")), exprs_values = "logcounts")
p14SCE <- calculateQCMetrics(p14SCE, feature_controls = list(ERCC =
isSpike(p14SCE,"ERCC"), MT = isSpike(p14SCE,"MT")))

#Keeping cells with more than 4000 transcript reads to ensure high quality cells are used
for analysis.
keepcell <- (p14SCE$total_counts >= 4000)
p14SCE <- p14SCE[,keepcell]
keep_gene <- rowSums(counts(p14SCE)>0)>0
p14SCE <- p14SCE[keep_gene,]


grep("^TOP2B",rownames(p14SCE))# index  4217

#Identify top2b negative cells

NoExpresTop <- counts(p14SCE)[4217,] == 0# 10811 cells have no top2b being
expressed
NoTop2b <- p14SCE[,NoExpresTop]

grep("^RHO$", rownames(NoTop2b))#index 14836
grep("^NRL$", rownames(NoTop2b)) #index 4534


topNnrlP <- NoTop2b[,counts(NoTop2b)[4534,]>0]
topNnrlPrhoP<- topNnrlP[,counts(topNnrlP)[14836,]>0]
#saveRDS(topNnrlPrhoP,"topNnrlPrhoP.rds")
topNnrlPrhoN <- topNnrlP[,counts(topNnrlP)[14836,]==0]
#saveRDS(topNnrlPrhoN,"topNnrlPrhoN.rds")

topNnrlN <- NoTop2b[,counts(NoTop2b)[4534,]==0]
topNnrlNrhoN <- topNnrlN[,counts(topNnrlN)[14836,]==0]
#saveRDS(topNnrlNrhoN,"topNnrlNrhoN.rds")
topNnrlNrhoP <- topNnrlN[,counts(topNnrlN)[14836,]>0]
#saveRDS(topNnrlNrhoP,"topNnrlNrhoP.rds")


#top2b expressing
grep("^TOP2B$",rownames(p14SCE)) #index 4232
range(counts(p14SCE)[4217,])# 0 to 28
```

```
highExpressingTop <- counts(p14SCE)[4217,] >=1
top2BSCE <- p14SCE[,highExpressingTop]
grep("^RHO$",rownames(top2BSCE)) #index 14836
grep("^NRL$",rownames(top2BSCE))#index 4534


topPnrlN <- top2BSCE[,counts(top2BSCE)[4534,] == 0]
topPnrlNrhoN <- topPnrlN[,counts(topPnrlN)[14836,] == 0]
#saveRDS(topPnrlNrhoN,"topPnrlNrhoN.rds")
topPnrlNrhoP <- topPnrlN[,counts(topPnrlN)[14836,] >= 1]
#saveRDS(topPnrlNrhoP,"topPnrlNrhoP.rds")
topPnrlP <- top2BSCE[,counts(top2BSCE)[4534,] >=1 ]
topPnrlPrhoP <- topPnrlP[,counts(topPnrlP)[14836,] >= 1]
#saveRDS(topPnrlPrhoP,"topPnrlPrhoP.rds")
topPnrlPrhoN <- topPnrlP[,counts(topPnrlP)[14836,]==0]
#saveRDS(topPnrlPrhoN,"topPnrlPrhoN.rds")
```

B.3 Finding Differentially Expressed Genes and Clustering

To find differentially expressed genes the method found in E. Macosko et al's paper "Highly Parallel Genome-wide expression profiling of individual cells using nanoliter droplets" was adapted. In this method, the variation across all cells for each gene was calculates as well as the dispersion measure. Dispersion measure is calculated by dividing the variation of each gene by the average log counts of the gene. Genes are then placed into twenty "similarity" bins based on their mean log expression. The z-score for each gene is then calculated by subtracting the average dispersion measure of all genes in a single bin from the dispersion measure of the gene and divided by the standard deviation of the dispersion measure across all genes in the bin. High variation genes are genes that have a z-score greater than 1.7. These genes, along with genes of interest for our purposed are kept. SC3 then calculates the number of clusters that can be found from the differentially expressed genes. The sc3 method places the cells into clusters and labels the

genes that are marker genes for each cluster. The following code is utilized in the command line. It is used for all groups of data.

```
library(scater)
library(SC3)
clusterFunction <- function(SCEObject,name){
  #find spike ins
  isSpike(SCEObject,"ERCC")<- grep("^ERCC-",rownames(SCEObject))
  isSpike(SCEObject,"MT") <- grep("^MT-", rownames(SCEObject))
  #renormalize the data for the group
  logcounts(SCEObject) <- log2(calculateCPM(SCEObject, use.size.factors = FALSE) +
1)
  SCEObject #23462 genes 5812 cells
  #need to rerun QCmetrics because different percentages will present themselves in each
group
  SCEObject <- calculateQCMetrics(SCEObject,feature_controls = list(ERCC =
isSpike(SCEObject,"ERCC"),MT = isSpike(SCEObject,"MT")), exprs_values =
"logcounts")
  SCEObject <- calculateQCMetrics(SCEObject,feature_controls = list(ERCC =
isSpike(SCEObject,"ERCC"),MT = isSpike(SCEObject,"MT")), exprs_values =
"counts")
  #remove  genes with no expression
  keepGene <- rowSums(logcounts(SCEObject)>0)>0
  SCEObject <- SCEObject[keepGene,]

  #this field is required to be able to run sc3 prepare
  rowData(SCEObject)$feature_symbol <- rownames(SCEObject)


  #calculate variation of logcounts per gene/feature
  variation <- apply(logcounts(SCEObject),1,var)
  #save in SCE rowData
  rowData(SCEObject)$varianceOfLogCount <- as.vector(variation)

  #calculate dispersion measure variance/mean of gene
  rowData(SCEObject)$dispersionMeasure <-
(rowData(SCEObject)[,"varianceOfLogCount"] /
rowData(SCEObject)[,"mean_logcounts"])

  #create similarity bins, each bin contains the indexes "Highly Parallel Genome-wide
expression profliling of individual cells using nanoliter droplets" by E. Macosko et al.
  bin1 <- which(rowData(SCEObject)[,"mean_logcounts"] >= 1)
  bin2 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .25 &
rowData(SCEObject)[,"mean_logcounts"] < 1)
```

```
  bin3 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .15 &
rowData(SCEObject)[,"mean_logcounts"] < 0.25)
  bin4 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .1 &
rowData(SCEObject)[,"mean_logcounts"] < 0.15)
  bin5 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .068 &
rowData(SCEObject)[,"mean_logcounts"] < 0.1)
  bin6 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .04 &
rowData(SCEObject)[,"mean_logcounts"] < .068)
  bin7 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .028 &
rowData(SCEObject)[,"mean_logcounts"] < .04)
  bin8 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .02 &
rowData(SCEObject)[,"mean_logcounts"] < .028)
  bin9 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .012 &
rowData(SCEObject)[,"mean_logcounts"] < .02)
  bin10 <-which(rowData(SCEObject)[,"mean_logcounts"] >= .007 &
rowData(SCEObject)[,"mean_logcounts"] < .012)
  bin11 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .0035 &
rowData(SCEObject)[,"mean_logcounts"] < .007)
  bin12 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .0018 &
rowData(SCEObject)[,"mean_logcounts"] < .0035)
  bin13 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .0009 &
rowData(SCEObject)[,"mean_logcounts"] < .0018)
  bin14 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .0004 &
rowData(SCEObject)[,"mean_logcounts"] < .0008)
  bin15 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .0003 &
rowData(SCEObject)[,"mean_logcounts"] < .0006)
  bin16 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .00015 &
rowData(SCEObject)[,"mean_logcounts"] < .0003)
  bin17 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .00008 &
rowData(SCEObject)[,"mean_logcounts"] < .00015)
  bin18 <-which(rowData(SCEObject)[,"mean_logcounts"] >= .000046 &
rowData(SCEObject)[,"mean_logcounts"] < .00008)
  bin19 <- which(rowData(SCEObject)[,"mean_logcounts"] >= .000015 &
rowData(SCEObject)[,"mean_logcounts"] < .000046)
  bin20 <- which(rowData(SCEObject)[,"mean_logcounts"] <.000015)

  #binlistnames and binlist variables are created for easier iteration and calculation of
zscores
  binlistnames <- c("bin1",
"bin2","bin3","bin4","bin5","bin6","bin7","bin8","bin9","bin10","bin11","bin12","bin13"
,"bin14","bin15","bin16","bin17","bin18","bin19","bin20")
  binlistVariables <-
list(bin1,bin2,bin3,bin4,bin5,bin6,bin7,bin8,bin9,bin10,bin11,bin12,bin13,bin14,bin15,bi
n16,bin17,bin18,bin19,bin20)
  count = 1
```

```
  #find zscore. divide each individual (gene dispersion measure subtract similarity group's
mean) by standard deviation of the similarity group
  for ( i in binlistnames){
    accessIndex <- binlistVariables[[count]]
    rowData(SCEObject)[accessIndex,"zscoreDispersion"] <-
(rowData(SCEObject)[accessIndex,"dispersionMeasure"]-
mean(rowData(SCEObject)[accessIndex,"dispersionMeasure"]))/
sd(rowData(SCEObject)[accessIndex,"dispersionMeasure"])
    count <- count + 1
  }

  #keep genes with a zscore higher than 1.7 and genes of interest
  #this is a logical of all high variation genes
  highVariationGenes <- rowData(SCEObject)[,"zscoreDispersion"] > 1.7

  #create logicals for genes of interest because this format can then be merged into the
high variation gene logical
  Top2b <- grepl("^TOP2B$", rownames(SCEObject))
  Crx <- grepl("^CRX$",rownames(SCEObject))
  Rho <- grepl("^RHO$",rownames(SCEObject))
  Opn <- grepl("^OPN1",rownames(SCEObject))
  Pde6 <- grepl("^PDE6", rownames(SCEObject))
  Rcv <- grepl("^RCVRN", rownames(SCEObject))
  Nrl <- grepl("^NRL", rownames(SCEObject))
  Tulp1 <- grepl("^TULP1$", rownames(SCEObject))
  Sag <-  grepl("^SAG$", rownames(SCEObject))
  Nr2e3 <- grepl("^NR2E3",rownames(SCEObject))
  Olig2 <- grepl("^OLIG2", rownames(SCEObject))
  Otx2 <- grepl("OTX2",rownames(SCEObject))
  Ascl1 <- grepl("ASCL1",rownames(SCEObject))
  Vsx2 <- grepl("VSX2",rownames(SCEObject))
  Rorb <- grepl("RORB",rownames(SCEObject))
  Prdm1 <- grepl("PRDM1",rownames(SCEObject))
  Rxrg <- grepl("RXRG", rownames(SCEObject))
  Glo1 <- grepl("GLO1",rownames(SCEObject))
  Smim13 <- grepl("SMIM13", rownames(SCEObject))
  Bc <- grepl("BC030499", rownames(SCEObject))
  Rp <- grepl("RPH3A",rownames(SCEObject))
  #keep genes of interest, high variation genes and spike ins
  #create a list of large logicals
  keeptheseGenes <- list(highVariationGenes,Top2b, Rho, Crx, Opn,
Pde6,Rcv,Nrl,Tulp1,Sag,Nr2e3,Olig2,Otx2,Ascl1,Vsx2,Rorb,Prdm1,Rxrg,Glo1,Smim13,
Bc, Rp,isSpike(SCEObject))
  keepGenes <- Reduce("|",keeptheseGenes) #merge all the logicals on the "OR" operator
  keepGenes[is.na(keepGenes)] <- FALSE #if NA set to false
  SCEObject <- SCEObject[keepGenes,]
```

```
  #get rid of genes that are not being expressed in logcounts
  SCEObject <- SCEObject[,SCEObject$total_logcounts >0]


  SCEObject <- sc3_prepare(SCEObject) #prepare the set for cluster estimation
  #estimate the number of clusters present in the data
  SCEObject <- sc3_estimate_k(SCEObject)
  kEstimation = metadata(SCEObject)$sc3$k_estimation

  #run tsne on data
  SCEObject<-runTSNE(check_duplicates = FALSE, SCEObject, perplexity = 30)
  fit <- kmeans(reducedDim(SCEObject,"TSNE"), kEstimation, nstart =40)
  colData(SCEObject)$clusterTSNE <- fit$cluster

  #this for some reason does not work in the command line call, so I ran this same code in
R for get the plots (officially)
  plotTSNE(SCEObject, colour_by = 'clusterTSNE')
  plotHighestExprs(SCEObject[!isSpike(SCEObject),],n=20)


  #find marker cells and consensus
  SCEObject <- sc3(SCEObject, ks = kEstimation, biology = TRUE)

  saveRDS(SCEObject, file = paste(name, "SC3.rds", sep=""))

  #save plots
  pdf(paste(name, "pdf", sep="Consensus."))
  sc3_plot_consensus(SCEObject,k = kEstimation)
  dev.off()
  png(paste(name, "png", sep="Conensus."))
  sc3_plot_consensus(SCEObject,k = kEstimation)
  dev.off()


}


SCEObject <- commandArgs(trailingOnly=TRUE)
# test if there is at least one argument: if not, return an error
if (length(SCEObject)==0) {
  stop("At least one argument must be supplied (input file).n", call.=FALSE)
} else if (length(SCEObject)==1) {
  #keep the name of the object so we can save graphs and new data
  name <- SCEObject
  name <- gsub(".rds","",name)
```

```
SCE <- readRDS(SCEObject)

clusterFunction(SCE,name)
```

B.4 Comparison of Highly Expressed Genes

A comparison of the top 100 genes in each group was done by inputting two groups into the comparison function. The comparison function utilized scater's rank object that is automatically generated. Using R's %in% function allowed us to create three files containing the genes that are shared between the groups and genes that are only in one of the groups. This code was made to be run in the command line and used over multiple groups.

```
library(scater)

comparison <- function(sce1,sce2,sce1Name,sce2Name){
  #get indices in order based on the highest expression without spike ins
  rank1 <-order(rowData(sce1)$rank_logcounts[!isSpike(sce1)],decreasing = TRUE)

  rank2 <-order(rowData(sce2)$rank_logcounts[!isSpike(sce2)],decreasing = TRUE)

  #rownames in order without spike in
  rownames1 <- rownames(sce1)[!isSpike(sce1)]
  rownames2 <- rownames(sce2)[!isSpike(sce2)]

  #get rownames of highest expression 1-50
  ranked1 <- rownames1[rank1[1:100]]
  ranked2 <- rownames2[rank2[1:100]]

  in1<-paste("In",sce1Name,sep = "_")
  #in 1 not in 2
  write.table(ranked1[!(ranked1 %in% ranked2)],file= paste(in1,"txt",sep = "."))

  in2 <- paste("In", sce2Name,sep="_")
  #in 2 not in 1
  write.table(ranked2[!(ranked2 %in% ranked1)],file =paste(in2,"txt",sep = ".") )

  #shared genes
  shared <-paste(sce1Name,sce2Name, sep="_")
  write.table(ranked1[ranked1 %in% ranked2],file = paste(shared,"txt",sep="."))
```

```
}


args <- commandArgs(trailingOnly=TRUE)
# test if there is at least one argument: if not, return an error
if (length(args)<2) {
  stop("At least two arguments must be supplied (input file).n", call.=FALSE)
} else if (length(args)==2) {
  sce1Name <- args[1]
  sce1Name <- gsub("/home/aep139/retina/allMyRData/SC3Complete/","",sce1Name)
  sce1Name <- gsub("SC3.rds","",sce1Name)
  sce1 <- readRDS(args[1])
  sce2Name <- args[2]
  sce2Name <- gsub("SC3.rds","",sce2Name)
  sce2Name <- gsub("/home/aep139/retina/allMyRData/SC3Complete/","",sce2Name)
  sce2 <- readRDS(args[2])
  comparison(sce1,sce2,sce1Name,sce2Name)
}
```

Appendix C

Supplementary Tables

C1 Table of Top 100 Genes Expressed in TOP2B(+)NRL/CRX(+)RHO(+)

Shared	NRL(+)	CRX(+)
"RHO"	"SYT11"	"SYT4"
"SAG"	"JUN"	"PRKCA"
"PDE6G"	"SPOCK3"	"LHX4"
"TULP1"	"CACNG4"	"LMO4"
"RCVRN"	"SPARCL1"	"TRNP1"
"PDE6B"	"DNER"	"TMEM215
"NRL"	"STMN2"	"FAM19A3"
"NR2E3"	"RPH3A"	"NNAT"
"MEG3"	"NSG2"	"RUNX1T1"
"CRX"		
"PDE6A"		
"OTX2"		
"TOP2B"		
"GNAO1"		
"CELF4"		
"RORB"		
"SNHG11"		
"GNGT2"		
"GNB3"		
"GRIA2"		
"NRXN3"		
"PCP4"		
"MARCKS"		
"GLO1"		
"PDE6H"		
"GUCA1A"		
"GNG3"		

"SCG2"
"TUBB2A"
"APP"
"NAP1L5"
"ITM2C"
"FOS"
"GPM6A"
"STMN3"
"LIN7A"
"BASP1"
"ATP1B1"
"PTPRD"
"EGR1"
"TCF4"
"GNG13"
"SPHKAP"
"HLF"
"GNAT2"
"CADPS"
"PDE6D"
"OPN1SW"
"TRPM1"
"GUCY1A3"
"NEUROD4"
"UCHL1"
"GLUL"
"GM4792"
"NDRG4"
"ARR3"
"SIX3"
"LRTM1"
"NRXN2"
"OPN1MW"

"PAX6"
"ISL1"
"GABRA1"
"CCDC136"
"2900011O08RIK"
"PCP2"
"TKT"
"PROX1"
"THSD7A"
"CABP5"
"CPLX2"
"SMIM13"
"TFAP2B"
"RAB3C"
"SCGN"
"ELAVL3"
"PDE6C"
"VSX2"
"KCNE2"
"DKK3"
"SLC6A1"
"SYNPR"
"CAR10"
"BC030499"
"ADARB1"
"GRM6"
"FRMD3"
"B3GALT2"
"SLC24A3"
"GPR179"
"SLIT2"

C2 Table Of TOP2B(+)NRL(+)RHO(+) Marker Genes for Clusters

1 (108 cells)	2 (183 cells)	3 (178 cells)	4 (7 cells)	5 (11 cells)	6 (37 cells)	7 (148 cells)
"PRDM1"	"GNGT2"	"PDE6A"	"FOS"	"SPARC"	"KCNE2"	"NEUROD4"
"ARR3"	"GLO1"	"CRX"	"SMIM13"	"SPC25"	"PDE6C"	"TMCC3"
"OPN1MW"	"GUCA1A"		"EPAS1"	"JUN"	"FAM19A3"	"NRXN3"
	"PDE6D"		"EGR1"		"GNB3"	"GM17750"
	"GNAT2"		"RORB"		"APOE"	"OTX2"
	"CNGB3"		"CDKN1C"			"CDH9"
	"PDE6H"		"JUNB"			"A330050F15RIK"
	"OPN1SW"		"NR2E3"			"TCF4"
	"CCDC136"					"CABP2"
						" "ST18"
						" "SULF2"
						" "LPHN2"
						" "EPHA7"
						" "GABRR2"
						" "TMEM215"
						" "FRMD3"
						" "NFIB"
						" "E130114P18RIK"
						" "NFIA"
						" "3110047P20RIK"
						" "ZFP804B"
						" "GSG1"
						" "PTPRZ1"
						" "FAM196A"
						" "CDH11"
						" "NYX"

8 (112 cells)	9 (95 cells)	10 (136 cells)	11 (95 cells)	12 (37 cells)
"GABRA1"	"LIN7A"	"PHLDA1"	"UTRN"	"MARCKS"
"GLRA1"	"IGF1"	"GRIK2"	"COL6A1"	"AKAP12"
"BC030499"	"GM4792"	"LRRN3"	"ADARB1"	"SLC35F1"
"SCGN"	"GRM6"	"TFAP2A"	"C1QL1"	"PCBD1"
"SLITRK6"	"SEBOX"	"SYNPR"	"CACNG4"	"NSG2"
"CADPS"	"CAR10"	"CELF4"	"CDC42EP4"	"TENM2"
"GRIK1"	"GPR179"	"EPB4.1L4A"	"KCNIP1"	"EBF1"
"LHX4"	"VSX2"	"ISOC1"	"COL23A1"	"NEFH"
"SCG2"	"ISL1"	"RGS8"	"GRIA1"	"6330403K07RIK"
"SPHKAP"	" "CACNA2D3"	"TFAP2B"	"HLF"	" "MEG3"
"LAMP5"	" "LRTM1"	"KCND3"	"RGS7BP"	" "VSNL1"
"OTOR"	" "AMIGO2"	"4833424O15RIK"	"DTNBP1"	" "RAB3C"
"VSX1"	" "PCP4"	"PTPRO"	"ID4"	" "HECW1"
"AI504432"	" "GNG13"	"LRRTM1"	"CPLX2"	" "TUBB2A"
"ZFHX4"	" "QPCT"	"FXYD6"	"CXCL14"	" "CARTPT"
"PTPRD"	" "CASP7"		"LHFPL2"	" "STMN4"
"SLIT2"	" "RNF152"		"GNG2"	" "NEFM"
"THSD7A"	" "B3GALT2"		"TKT"	" "NEFL"
"PDE1C"	" "TGFB2"		"CPNE6"	" "BASP1"
"LYVE1"	" "PROX1"		"PCDH17"	" "NCALD"
"TNNT1"	" "ITM2C"		"CACNG2"	" "BAI1"
"CDH8"	" "BHLHE23"		"AI848285"	" "LYNX1"
"ESAM"	" "DAPL1"		"FILIP1L"	" "LY6E"
"A730046J19RIK"	" "NTNG1"		"HUNK"	" "RBFOX2"
	" "DAB1"		"CPNE5"	" "2900011O08RIK"
	" "TRNP1"		"DLGAP1"	" "FGF12"
	" "CAR8"		"SIX3OS1"	" "LSAMP"
	" "FAM184B"		"SIX3"	" "GAP43"
	" "PRDM8"		"CAMK4"	" "ALCAM"
	" "CNTN4"		"HBEGF"	" "RBFOX1"
	" "STRIP2"		"SYT7"	" "NCAM2"
	" "NDNF"		"GFRA1"	" "APP"
	" "CRYM"		"NRXN2"	" "NRXN1"
	" "CABP5"		"CCDC88B"	" "SEMA6A"
	" "VSTM2B"		"C1QL2"	" "TRPM3"
	" "TRPM1"		"GLUL"	" "SORCS1"
	" "RLBP1"		"SERPINE2"	" "FLRT1"
	" "GRM5"		"ARL4C"	" "GNG3"
	" "PCP2"		"PAX6"	" "ATP1B1"
	" "MT2"		"FRMD5"	" "RIMS1"
			"SLC32A1"	" "MYO1B"
			"GAD2"	" "RESP18"
			"RND3"	" "SLC4A3"
			"GAD1"	" "DNER"

11 (95 cells)	12 (37 cells)	13 (159 cells)
"NHLH2"	" "BDNF"	"LGR5"
"FNBP1L"	" "PLCB1"	"VSTM2A"
"NPNT"	" "SLC24A3"	"GAS1"
"DDAH1"	" "NRSN2"	"FBXO32"
"BHLHE22"	" "NNAT"	"DUSP1"
"KCNAB1"	" "VSTM2L"	"PRDX6"
"GRIA2"	" "SNHG11"	"VIM"
"GUCY1A3"	" "TSHZ2"	"PTPRT"
"PDE4B"	" "PHACTR3"	"NEBL"
"NECAB1"	" "STMN3"	" "ZEB2"
"FUT9"	" "SCN3A"	" "ZFP804A"
"WBSCR17"	" "SCN2A1"	" "SLC6A9"
"NPTX2"	" "SCN1A"	" "TSC22D4"
"CACNA2D1"	" "PCDH10"	" "SPON1"
"PCDH7"	" "STMN2"	" "NFIX"
"SLC6A1"	" "MAB21L2"	" "MT1"
"KCNA1"	" "SYT11"	" "CAMKV"
"KCNA6"	" "TPM3"	" "ARHGAP20"
"PTN"	" "ELAVL4"	
"DKK3"	" "PTPRF"	
"FOSB"	" "HPCA"	
"ZMAT4"	" "RUNX1T1"	
"GPM6A"	" "SPARCL1"	
"SPOCK3"	" "CPLX1"	
"LPL"	" "SEZ6L"	
"CACNA2D2"	" "RPH3A"	
"ELAVL3"	" "RELN"	
"CRABP1"	" "CDK14"	
"UACA"	" "KCNIP4"	
"UNC13C"	" "UCHL1"	
"CNKSR2"	" "GABRA2"	
"GABRA3"	" "DYNC1I1"	
	" "NAP1L5"	
	" "SNCA"	
	" "MGLL"	
	" "GPR123"	
	" "ZFHX3"	
	" "CALB2"	
	" "ITGB1"	
	" "CHRNB3"	
	" "CHRNA6"	
	" "NETO2"	
	" "GNAO1"	
	" "MT3"	
	" "NDRG4"	
	" "UBASH3B"	
	" "THY1"	
	" "AI593442"	
	" "PCDH19"	
	" "GRIA3"	
	" "L1CAM"	

C3 Table of 100 Genes Expressed in Isolated TOP2B(+)NRL(+)CRX(-)RHO(+) and TOP2B(+)NRL(-)CRX(+)RHO(+)

CRX(+)NRL(-)	CRX(-)NRL(+)	Shared
"SYT4"	"SYT11"	"MEG3"
"GNG13"	"RCVRN"	"GNAO1"
"TRPM1"	"NRL"	"GRIA2"
"NEUROD4"	"PDE6B"	"CELF4"
"ISL1"	"GAP43"	"SNHG11"
"PCP2"	"C1QL1"	"NRXN3"
"CRX"	"RPH3A"	"TUBB2A"
"LRTM1"	"NR2E3"	"PCP4"
"GNB3"	"TPM3"	"SCG2"
"CABP5"	"NEFL"	" "APP"
"SCGN"	"CACNG4"	" "MARCKS"
"VSX2"	"CALB2"	" "GNG3"
"FRMD3"	"CDK14"	" "NAP1L5"
"CAR10"	"THY1"	" "CADPS"
"B3GALT2"	"NEFM"	" "RHO"
"GPR179"	"ELAVL4"	" "LIN7A"
"LHX4"	"GRIA3"	" "PTPRD"
"GM4792"	"NRXN1"	" "GUCY1A3"
"NNAT"	"FRMD5"	" "STMN3"
"GRM6"	"RBFOX1"	" "ATP1B1"
"LMO4"	"RBFOX2"	" "TOP2B"
"ZFHX4"	"PDE6A"	" "HLF"
"ZFHX3"	"SLC4A3"	" "BASP1"
"SLC24A3"	"STMN4"	" "ITM2C"
"BC030499"	"PTN"	" "SAG"
"TRNP1"	"ID4"	" "GPM6A"
"CACNA2D1"	"SYT7"	" "UCHL1"
"SLIT2"	"AI593442"	" "PROX1"

"GRIK1"	"GLO1"	" "OTX2"
"TMEM215"	"NCALD"	" "NRXN2"
"GNGT2"	"GAD2"	" "NDRG4"
"IGF1"	"SCN2A1"	" "GABRA1"
"LPHN2"	"6330403K07RIK"	" "SPHKAP"
		" "THSD7A"
		" "TCF4"
		" "2900011O08RIK"
		" "PAX6"
		" "RAB3C"
		" "SLC6A1"
		" "TKT"
		" "GLUL"
		" "SIX3"
		" "RORB"
		" "ELAVL3"
		" "CPLX2"
		" "SPOCK3"
		" "LSAMP"
		" "SPARCL1"
		" "DNER"
		" "TFAP2B"
		" "RUNX1T1"
		" "DLGAP1"
		" "FOS"
		" "PDE6G"
		" "ZEB2"
		" "ADARB1"
		" "TULP1"
		" "SYNPR"
		" "GAD1"
		" "NSG2"
		" "DKK3"

		" "GNG2"
		" "STMN2"
		" "CACNA2D2"
		" "VSNL1"
		" "SLC32A1"
		" "DTNBP1"

References

1. Haque, Ashraful et al. “A practical guide to single-cell RNA sequencing for biomedical research and clinical applications.” *Genome Medicine.* 9.75 (2017).
2. Shekhar, Karthik et al. “Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics.” *Cell*. 166.5 (2016): 1308-1323e30.
3. Jaiten, Diego Adhemar et al. “Massively Parallel Single-Cell RNA-Seq for Marker Free Decomposition of Tissues in Cell Types.” *Science*. 343.6172 (2014): 776-779.
4. Zeisel, Amit et al. “Cell types in the mouse cortex and hippocampus revealed by single cell RNA-seq.” *Science*. 347.6226 (2015): 1138-1142.
5. Li, Ying et al. “Top2b is Involved in the Formations of Outer Segment and Synapse During Late-stage Photoreceptor Differentiation by Controlling Key Genes of Photoreceptor Transcriptional Regulatory Network.”*Journal of Neuroscience Research.* 95 (2017): 1951-1964.
6. Macosko, Evan Z. et al. “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.” *Cell.* 161:5 (2015):1202-1214.
7. Bollimpalli, V. Satish et al. “Topoisomerase IIB and its role in different biological contexts.” *Archives of Biochemistry and Biophysics.* 633 (2017): 78-84.
8. Daniszewski, Maciej et al. “Single Cell RNA Sequencing of Stem Cell-Derived Retinal Ganglion Cells.” *Scientific Data.* 5 (2018): 180013.
9. Eon, Chang-Jin et al. “The Major Cell Populations of the Mouse Retina.” *Journal of Neuroscience.* 18.21 (1998): 8936-8946.
10. Hao, Hong et al. “Transcriptional Regulation of Rod Photoreceptor Homeostasis Revealed by *In Vivo* NRL Targetome Analysis.” *PLOS Genetics*. (2012)
11. McCarthy DJ, Campbell KR, Lun ATL and Wills QF “Scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R.” *Bioinformatics*, (2017).
12. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR and Hemberg M. “SC3 - consensus clustering of single-cell RNA-Seq data.” *Nature Methods.* (2017).
13. Brzezinskin, JA et al. “Blimp1 (Prdm1) prevents re-specification of photoreceptors into retinal bipolar cells by restricting competence.” *Dev Bio.* 384.1 (2013): 194-204.
14. Aligianis, IA et al. “Mapping of novel locus for achromatopsia (ACHM4) to 1p and identification of germline mutation in the alpha subunit of cone transducin (GNAT2).” *J Med Genet.* 39.9 (2002): 656-660.
15. Kohl, S. et al. “A nonsense mutation in PDE6H causes autosomal-recessive incomplete achromatopsia.” *Am J Hum Genet*. 91.3 (2012):527-32.
16. Hirano, A. A. et al. “Synaptotagmin-4 Expression in Mouse Horizontal Cells.” *Investigative Ophthalmology & Visual Science.* 46.2798 (2007).
17. Ruether, K. et al “PKCA is Essential for the Proper Activation and Termination of Rod Bipolar Cell Response.” *Invest Opthalmol Vis Sci.* 51.11(2010):6051-6058.

18. Duquette, P.M. et al. “Loss of LMO4 in the Retina Leads to Reduction of GABAergic Amacrine Cells and Functional Deficits.” *PLOS ONE* 5.10 (2010): e13232
19. Balasubramanian, Revathi et al. “Expression of LIM-Homeodomain Transcription Factors in the Developing and Mature Mouse Retina.” *Gene expression patterns : GEP* 14.1 (2014): 1–8.
20. Cheng H, Aleman TS, Cideciyan AV, Khanna R et al. “In vivo function of the orphan nuclear receptor NR2E3 in establishing photoreceptor identity during mammalian retinal development.” *Hum Mol Genet*. 15.17 (2006):2588-602.
21. Shaw, G. et al. “Transcriptional regulation of synaptotagmin 11 in retinal ganglion cells.” *Invest Opthalmol Vis Sci.*51.15 (2013):6205.
22. Mallo, G. et al. “Expresision of c-fos and c-jun in rat retina following protracted illumination.” *Brain Res*. 693.1(1995): 196-200.
23. Guo, Ying et al. “Retinal Cell Responses to Elevated Intraocular Pressure: A Gene Array Comparison between the Whole Retina and Retinal Ganglion Cell Layer.” *Investigative Ophthalmology & Visual Science* 51.6 (2010): 3003–3018.
24. Chung, Woosung et al. “Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer.” *Nature Comm.* 8:15081 (2017).