ABSTRACT

Title of Dissertation: ANALYTICAL APPROACHES FOR COMPLEX MULTI-BATCH -OMICS DATASETS AND THEIR APPLICATION TO NEURONAL DEVELOPMENT

Theresa Ann Alexander, Doctor of Philosophy, 2023

Dissertation directed by: Assistant Professor Colenso M. Speer, Department of Biology
Professor Najib M. El-Sayed, Department of Cell Biology and Molecular Genetics

High-throughput sequencing methods are extremely powerful tools to quantify gene expression in bulk tissue and individual cells. Experimental designs often aim to quantify expression shifts to characterize developmental trajectories, disease states, or cellular drug responses. Experimental and genetic methods are also rapidly evolving to capture specific aspects of gene expression such as in targeting individual cell types, regulatory stages, and spatially resolved cell subcompartments. These studies frequently involve a variety of experimental conditions that require many samples to guarantee sufficient statistical power for subsequent analyses. These studies are frequently processed in multiple batches due to limitations on the number of samples

that can be collected, processed, and sequenced at once. To eliminate erroneous results in subsequent analyses, it is necessary to deconvolve non-biological variation (batch effect) from biological signal. Here, we explored variational contributions in multi-batch high throughput sequencing experiments by developing new methods, evaluating heterogeneity-contributors in an axon-TRAP-RiboTag protocol case-study, and highlighting biological results from this protocol.

First, we describe iDA, a novel dimensionality reduction method for high-throughput sequencing data. High-dimensional data in complex, multi-batch experiments often result in discrete clustering of samples or cells. Existing unsupervised linear dimensionality reduction methods like PCA often do not resolve discreteness simply with projections of maximum variance. We show that iDA can produce better projections for separating discrete clustering that correlates with known experimental biological and batch factors.

Second, we provide a case study of special considerations for a complex, multi-batch high throughput experiment. We investigated the multi-faceted heterogenic contributions of a study using the axon-TRAP-RiboTag translatomic isolation protocol in a neuronal cell type. We show that popular batch-correction methods may reduce signal due to true biological heterogeneity in addition to technical noise. We offer metrics to help identify biological signal-driven batch-differences.

Lastly, we employ our understanding of variational contributions in the intrinsically photosensitive retinal ganglion cell (ipRGC) -omics case study to explore the biological transcriptomic and translatomic coordination. Our analysis revealed ipRGCs participate in

subcompartment-specific local protein translation. Genetic perturbations of photopigment-driven neuronal activity led to global tissue transcriptomic shifts in both the retina and brain targets, but the ipRGC axonal-specific translatome was unaltered.

ANALYTICAL APPROACHES FOR COMPLEX MULTI-BATCH -OMICS DATASETS AND THEIR APPLICATION TO NEURONAL DEVELOPMENT

by

Theresa Ann Alexander

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:

Dr. Colenso M. Speer, Co-Chair
Dr. Najib M El-Sayed, Co-Chair
Dr. Joshua Singer
Dr. Michael P. Cummings
Dr. Robert Patro, Dean's Representative

Preface

The work in this dissertation was conducted over the span of two research labs. The first part of my dissertation work was spent working with Dr. Héctor Corrada Bravo. Chapter 1 introduces the work conducted in the Corrada Bravo group. The algorithm adapted as a result of the work with Dr. Corrada Bravo is Chapter 2 of this dissertation which are peer-reviewed published works. After Dr. Corrada Bravo left the University of Maryland, I transitioned to co-mentorship under the advisors Dr. Colenso M. Speer and Dr. Najib M. El-Sayed. The second half of my dissertation work is presented in Chapters 3, 4, and 5. At the time of this writing, Chapter 4 is currently under preparation for submission and Chapter 5 is a portion of a paper which will be submitted within the calendar year pending integration with data which are currently being collected. For these 2 chapters, all the experimental data collection was performed by Rashmi Gupta and Shyrice M. Mitchell and all the data analysis was performed by Theresa Alexander. I would like to thank all the interdisciplinary co-author contributions to make this work impactful and overarching.

Chapter 2:

- Theresa A Alexander, Rafael A Irizarry, Héctor Corrada Bravo, Capturing discrete latent structures: choose LDs over PCs, *Biostatistics*, Volume 24, Issue 1, January 2023, Pages 1–16, https://doi.org/10.1093/biostatistics/kxab030. My individual contributions include: 1) literature review, 2) algorithm development,

3) experimental design, 4) algorithm assessment, 5) writing the manuscript, and 6) manuscript revisions.

Chapter 4:

- Theresa A Alexander, Rashmi Gupta, Ashton T. Belew, Shyrice M. Mitchell, Najib M. El-Sayed, Colenso M. Speer. “Validating the efficacy of Axon-TRAP RNA-Sequencing Results in Multi-Batch Studies”. This work will be submitted eNeuro in 2022. My individual contributions include: 1) literature review, 2) data analysis, 3) experimental design for “forensic bioinformatic” analysis, and 4) writing the manuscript.

Chapter 5:

- Theresa A Alexander*, Rashmi Gupta*, Ashton T. Belew, Shyrice M. Mitchell, Najib M. El-Sayed, Colenso M. Speer. “The effects of melanopsin on ipRGC cell autonomous and cell non-autonomous programming”. This work will be submitted to a peer-reviewed journal in 2022 in conjunction with on-going work. My individual contributions include: 1) literature review, 2) data pre-processing, 3) data analysis, and 4) writing this portion of the manuscript.

Other publications in which I have completed or contributed to during my PhD work include:

- Alexander, T.A., Machiela, M.J. LDpop: an interactive online tool to calculate and visualize geographic LD patterns. *BMC Bioinformatics* 21, 14 (2020). https://doi.org/10.1186/s12859-020-3340-1.
- Gomez MA, Belew AT, Navas A, Rosales-Chilama M, Murillo J, Dillon LAL, Alexander TA, Martinez- Valencia A and El-Sayed NM (2021). Early Leukocyte Responses in Ex-Vivo Models of Healing and Non- Healing Human Leishmania (Viannia) panamensis Infections. Front. Cell. Infect. Microbiol. 11:687607. doi: 10.3389/ fcimb.2021.687607.
- Sam B. Choi*, Theresa A. Alexander*, Tarlan Vatan*, Chenghang Zhang, Shyrice M. Mitchell, Colenso M. Speer, Peter Nemes. Ultrasensitive Capillary Electrophoresis Trapped Ion Mobility Time-of-Flight Mass Spectrometry Reveals a Proteome Transition During Development of the Brain's Circadian Pacemaker. Submitted to Mol. Cell. Prot. December, 2022.
- Li CY, Bowers JM, Alexander TA, Lawrence K, Coyne JM, Li Y, Juntti SA. Olfactory receptor for female reproductive pheromonal signaling in an African cichlid. In Preparation.
- Gómez,M.A.*, Belew, A.T. , Vargas, D.A., Rebellon, D., Rosales-Chilama, M., Alexander, T.A., Parra, L.F.G, Saravia, N, El-Sayed, N.M.* (2022). Innate biosignatures of treatment outcome in patients with cutaneous leishmaniasis. In Preparation.

- Fernandez, O., Gómez,M.A.*, Belew, A.T., Rosales-Chilama, M., Alexander, T.A., El-Sayed, N.M., Saravia, N. (2021). Genotypic and transcriptomic profiling of virulent and avirulent strains of L. (V.) panamensis. In Preparation.
- Gómez,M.A.*, Belew, A.T., Fernandez, O., Rosales-Chilama, M., Alexander, T.A., Saravia, N., El-Sayed, N.M. * (2022). Human macrophage response and intrinsic resistance of L. (V.) panamensis to pentavalent antimony. In. Preparation.

Dedication

For my family and friends, two legged and four, who have walked this journey with me.

Acknowledgements

I am grateful for the many mentors, advisors, and friends who have helped and supported me through my doctoral journey.

Firstly, I would like to thank my family for all their support over the past 5 years. Louis, Mom, Dad, Colleen, and Ricky: thanks for celebrating my wins and cushioning my losses. I am so fortunate to have gone through this process with such a supportive group behind me.

I will always be grateful that my first introduction to my doctoral journey was with Dr. Héctor Corrada Bravo's group. During my initial time at UMD, Héctor gave me the independence and the confidence I needed to find myself as a student and researcher. Along with Héctor, I feel indebted to the Corrada Bravo lab members: Jay, Dom, Mohamed, and Aya, for setting the tone for a work environment I will now forever chase after.

Dr. Najib El-Sayed and Dr. Colenso Speer took me in as an orphaned graduate student and made the transition feel seamless even in the midst of a global pandemic. I am so thankful for the mentorship and support each of them has given me over the past 3 years. The skills I have acquired for interdisciplinary research and communication would not be if not for Najib and Colenso.

I am also grateful for members of my thesis committee who each offered guidance towards my growth as a researcher as well as navigating my extremely nonlinear doctoral journey (in alphabetical order): Dr. Michael Cummings, Dr. Rob Patro, and Dr. Josh Singer. Each of you have challenged and supported me while making me feel like a valuable member of the scientific community.

I also am greatly appreciative for all the mentorship, guidance, support, and friendship from Dr. Ashton Trey Belew. I learned so much from getting to work with Trey but most importantly was the importance of fun in research.

I would like to thank the members of the Speer and Singer labs for their support, help, and friendship throughout my work. I would specifically like to thank (in alphabetical order): Jackie Minehart, Morgan Musgrove, Andrea Orpia, Kat Pizano, Tarlan Vatan.

I would also like to send my love and thanks to members of the BISI and CBCB community, specifically Alexis Boleda, Jason Fan, and Ashley Della Fera. Without the many hours of talks for research and otherwise, I would not be where I am today.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

scRNA-seq	Single cell RNA-seq
RNA-seq	Bulk RNA-sequencing
GWAS/TWAS	Genome/transcriptome-wide association studies
SVA	Surrogate variable analysis
PBMC	Peripheral blood mononuclear cells
PCA	Principal component analysis
PCs	Principal components
t-SNE	t-Distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
iDA	iterative Discriminant Analysis
LDA	Linear discriminant analysis
LD	Linear discriminant
QDA	Quadratic discriminant analysis
TSS	Total sum of squares
RSS	Residual sum of squares
DI	Dunn Index
NK	Natural killer
MAF	Minor allele frequency
EUR	European
YRI	Yoruban
SNP	Single nucleotide polymorphism
eQTL	Expression quantitative trait loci
ERCC	External RNA Controls Consortium

GEMs	Gel beads in emulsion
ARI	Adjusted rand index
FDR	False discovery rate
RGCs	Retinal ganglion cells
cRGCs	Conventional retinal ganglion cells
ipRGCs	intrinsically photosensitive retinal ganglion cells
SCN	Suprachiasmatic Nucleus
dLGN	dorsal Lateral Geniculate Nucleus
P0,P8, P15,P60	Postnatal day 0,8,15,60
E20	Embryonic day 20
TRAP	Translating ribosome affinity purification
DE	Differential expression
GSEA	Gene set enrichment analysis
Opn4	Melanopsin
SC	Superior colliculus
bp	Base pairs
BBI-ACGT	Brain & Behavior Institute-Advanced Genomic Technologies Core
UMIs	Unique molecular identifiers
Het	Heterozygote
KO	Knock-out
CPM	Counts per million
log2FC	Log_2 Fold Change

Eph	Ephrin receptor
Efn	Ephrin ligand
GO	Gene ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
VEGF	Vascular endothelial growth factor
PDGF	Platelet-derived growth factor
HIF-1a	Hypoxia-induced factor 1a
RTKs	Receptor tyrosine kinases
DAG	Diacylglycerol
BCL	Binary base call
IGV	Integrative Genome Viewer

Chapter 1: Introduction to investigating sources of variational contributions in RNA-seq data

1.1 Variability in high dimensional data

The number of studies that rely on high dimensional biological data collection continues to grow. This trend has been intensified recently by the advent of single cell RNA-seq (scRNA-seq), which simultaneously measures genome-wide expression profiles for tens of thousands of cells. In these studies, dimensionality reduction techniques are commonly used to capture underlying discrete structure of the data in a small number of dimensions. This structure can be imposed by sources of variation that could be strategic (experimental), biological but unknown and potentially interesting, or unwanted and unintentionally introduced by laboratory techniques. In high dimensional transcriptomic and translatomic datatypes, accurate dimensionality reduction techniques are essential for visualizing the data structure in exploratory analyses, identifying any potential variation attributed to noise in the data, and cutting computational costs for downstream analysis such as clustering. Low-dimensional representations need to capture the underlying structure of the samples from true separable differences between study groups as well as potential unwanted batch variation. In the case of scRNA-seq, investigators are often interested in discovering previously unknown cell types based on their gene expression profile. In bulk RNA-sequencing (RNA-seq) investigators commonly vary experimental conditions, introducing batch effects that induce unwanted variability which may cause spurious associations in downstream analysis if not corrected for (Leek et al., 2010).

Whether heterogeneity across classes arises from unwanted sources of variation, variation of response to an experimental intervention, or different cell types in scRNA-seq experiments, it is vital to be able to identify a low dimensional embedding which describes the separation of classes. This embedding can be used as an interpretable reduction to characterize and investigate features which drive class separation or to correct for unwanted sources of variation in downstream analysis. Varying experimental conditions commonly introduce batch effects that induce unwanted variability (Leek et al., 2010). For example, the Geuvadis Project expression data (Lappalainen et al., 2013) shows clear evidence of batch effect, where the laboratory in which the samples were processed contributes to a large part of variation in the data. If unaccounted for, this batch effect may cause spurious associations in downstream analysis. This is particularly problematic when the outcome of interest is confounded with the variable responsible for the batch effect (Leek et al., 2010).

1.1.1 Genome/Transcriptome-wide association studies

In genome or transcriptome-wide association studies (GWAS/TWAS), where each locus or transcript is tested for association with an outcome, standard statistical tests assume observations are independent. When there is underlying structure in the data, genetic population structure for example, the independence assumption does not hold (Brown et al., 2018). To correct for this latent structure, state-of-the-art methods include population structure as a covariate in the model. In GWAS, self-reported ancestry is not reliable (Mersha & Abebe, 2015), and the latent structure attributed to

ancestry must be estimated. Similarly, in TWAS, sources of technical variability are impossible to predict and methods that estimate these unknown factors, such surrogate variable analysis (SVA) (Leek & Storey, 2007) have been proposed.

1.1.2 Single cell RNA-sequencing variability

A common motivation of studies using scRNA-seq is to detect novel cell-types based on the natural grouping of the transcriptomic profiles and identify marker genes that define these cell types (Jain, 2010). The current standard approach is to define cell-types using unsupervised clustering followed by statistical tests used to identify genes with greater expression in one cell type compared to the others. While this may give researchers valuable information for a particular gene of interest, the method fails to provide a broad summary of a set of genes which define cell type separation. A more useful approach would be to find an embedding space which defines the latent structure of the separation of cell types. Often in scRNA-seq when trying to identify putative cell types, the boundaries between them are ill-defined, making it difficult to define definitive cell type boundaries (Kiselev et al., 2019). Defining the latent structure which captures the features which best separate cell types in low dimensional space from each other would be preferred.

1.2 Linear dimensionality reduction methods

1.2.1 Unsupervised: Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is commonly used as a dimensionality reduction method in these applications since it finds principal components (PCs) that explain variation in the data. For sources of variation such as batch effects or genetic

ancestry the variation in the data set is largely attributed to factors not associated with the outcome variable and PCA is used to find an embedding to estimate and account for these latent factors. In other data types like scRNA-seq, PCA is used to reduce the computational cost of downstream analyses like clustering. The application of PCA to detect latent discrete structure assumes that latent classes will in fact account for most of the dataset variance. However, since PCA is unsupervised, it does not necessarily separate classes (Lever et al., 2017). When the clusters formed by latent classes fall in directions of maximum variance, PCA will indeed find an embedding where clusters separate, but if these clusters do not fall in directions of maximum variance, embeddings computed by PCA will fail to separate these clusters.

Since PCA seeks to find transformations that explain most variance in the data, it inherently favors separating clusters with large variance. A well-separated cluster with low variance will likely not be separated by the top embedding directions since it does not contribute much to the total variance of the data set. For data sets with an innate discrete latent structure containing clusters with unbalanced variances an alternative method for recognizing latent variables is needed.

1.2.1 Supervised: Linear Discriminant Analysis (LDA)

Another method used for dimensionality reduction, classification, and data visualization is linear discriminant analysis (LDA). LDA, like PCA, is a linear reduction method but class labels are used in order to find optimal separating

boundaries between them. The class labels are used to maximize the ratio of between class variation to within class variation. In instances where the major sources of variation within a dataset are known (such as experimental variables and batches), LDA produces weighted vectors which describe the data separation and variability. The contribution of each feature in the separability per dimension is interpretable since they are linear.

1.3 Non-Linear dimensionality reduction methods

Visualization techniques, such as t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP), use a different approach to retain local structure in the original high dimensionality data in the low dimensional embedding (McInnes et al., 2018; van der Maaten & Hinton, 2008). tSNE first computes joint probabilities of the pairwise similarity scores in the full dimension data space and the low-dimensional projection space. Next, the algorithm minimizes the Kullback-Leibler divergence between the joint probabilities and iterates until convergence. This method preserves small pairwise distances/local similarities between samples, but because the cost function is not convex, convergence to the global minima is not guaranteed. Similarly to tSNE, UMAP is also a neighborhood-preserving non-linear method, but offers improvements on computational complexity compared to tSNE. These methods often yield visualizations that capture latent discrete structure in data well. These methods are computationally expensive and in practice are usually used on data that is already of smaller dimension by using PCA as a first step, thus inheriting the issues raised about

PCA above. More importantly for the context of biological interpretation, embeddings obtained by t-SNE and UMAP are nonlinear, making it hard to interpret feature weights to determine those that may be contributing to the signal separating latent discrete structure in the data.

1.4 Contributions

For datasets with an innate discrete latent structure containing clusters with unbalanced variances an alternative method for recognizing latent variables is needed. If the experimental design is such that one may expect unequal variances between groups of interest, the resulting PCA + tSNE or UMAP visual representation will not accurately separable differences between distinct groups with small variance between them. In particular, in instances where a large portion of the variation between study groups is attributed to one group's large variation, the difference between the groups with smaller distances between them becomes unobservable. In chapter 2, we develop a novel dimensionality reduction method, iterative discriminant analysis (iDA). iDA as an iterative extension of linear discriminant analysis (LDA). This new proposed method will mitigate this bias and preserve the distinct separation between low variance groups.

Chapter 2: A Novel Dimensionality Reduction Method for high-throughput RNA-sequencing data

2.1 Introduction

In this chapter, we propose iterative Discriminant Analysis (iDA), an unsupervised dimensionality reduction technique to address the problems of PCA and nonlinear methods like t-SNE and UMAP discussed in chapter 1. iDA uses linear transformations to produce an embedding space that contains discriminatory information that optimally separates latent clusters. This method leverages linear discriminant analysis (LDA) in order to find transformations which both separate classes from each other while also minimizing variance within each class. This method improves on PCA as it defines the latent space while maintaining unobserved clustered architecture.

Denote the observed $N \times p$ data matrix as $\mathbf{X}$. LDA, like PCA, is a dimensionality reduction technique since it finds a linear projection $\mathbf{W}$ of the data features to best fit the data in a lower dimension producing an embedding $\mathbf{Y} = \mathbf{XW}$. A major difference between these two linear projection techniques is that LDA is a supervised method, while PCA is unsupervised. Unsupervised PCA finds projections which maximize the total variance in the data, which may or may not separate classes in these projections (Figure 2.1). If the class assignments are known a priori, LDA can be used to find projections $\mathbf{W}_{\text{LDA}}$ which best separate the classes. In Figure 2.1, the direction which maximizes variation is uninteresting since it does not separate the latent clustering. However, the LDA projection finds a distinct boundary between the

classes. Such as in this example, a direction with low variation, but potentially valuable, class defining information, is not guaranteed to be preserved in the first few principal components. When determining the number of PCs to use, it is common practice to evaluate the eigenvalues for the top PCs in an elbow plot to determine the best cutoff for which eigenvectors to include in the reduced embedding. Since the eigenvectors indicate the amount of variation attributed to its respective eigenvector, the top PCs which capture the majority of the variation in the data are commonly used, meaning the lowly variable but well defined clusters will not be preserved in the PCA embedding.

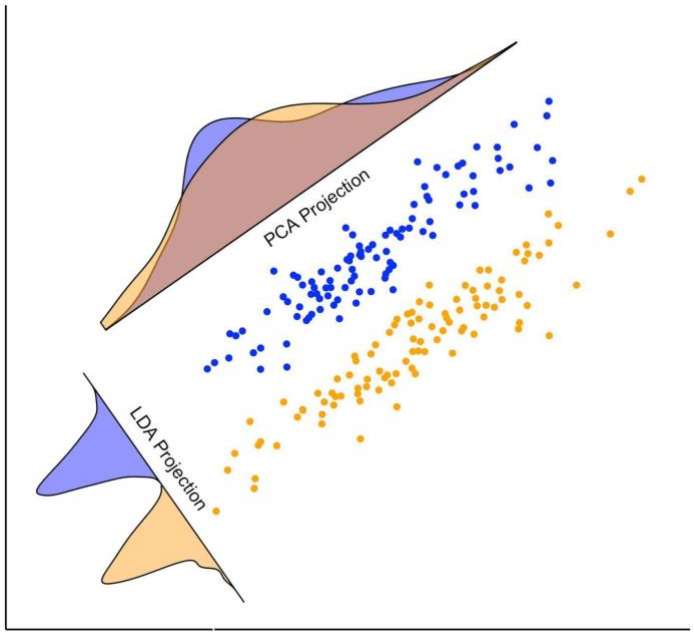


Figure 2.1. PCA versus LDA projection for data where data is clustered.
The PCA projection fails to separate the classes because they do not fall on the projection which maximizes variation for the data. The supervised LDA projection maximizes the ratio of between to within cluster variation.

PCA maximizes total variance of the embedded data by maximizing the objective function

$$J_{\mathrm{PCA}}(\mathbf{W};\mathbf{X}) = \|\tilde{\mathbf{S}}\|_2^2 = \|\mathbf{W}^T\mathbf{S}\mathbf{W}\|_2^2 \tag{2.1}$$

where $\tilde{\mathbf{S}}$ is the empirical covariance of the embedded data $\mathbf{Y}$

$$\tilde{\mathbf{S}} = \sum_i (\mathbf{y}_i - \tilde{\mu})(\mathbf{y}_i - \tilde{\mu})^T, \tag{2.2}$$

$$\text{with } \tilde{\mu} = \frac{1}{N}\sum_i \mathbf{y}_i. \tag{2.3}$$

and $\mathbf{y}_i$ is the ith row of embedding $\mathbf{Y}$ corresponding to the low-dimensional representation of the ith data observation ($\mathbf{x}_i$).

LDA uses observed assignments c_i to one of K classes to find projections which separate classes by maximizing the objective function:

$$J_{\mathrm{LDA}}(\mathbf{W};\mathbf{X}) = \frac{\|\tilde{\mathbf{S}}_{\mathbf{B}}\|_2^2}{\|\tilde{\mathbf{S}}_{\mathbf{W}}\|_2^2} = \frac{\|\mathbf{W}^T\mathbf{S}_{\mathbf{B}}\mathbf{W}\|_2^2}{\|\mathbf{W}^T\mathbf{S}_{\mathbf{W}}\mathbf{W}\|_2^2}, \tag{2.4}$$

where $\mathbf{S}_{\mathbf{B}}$ is the between class empirical covariance and $\mathbf{S}_{\mathbf{W}}$ is the within class empirical covariance matrix:

$$\tilde{\mathbf{S}}_{\mathbf{W}} = \sum_{k=1}^{K} \tilde{\mathbf{S}}_{\mathbf{k}} = \sum_{k=1}^{K}\sum_{c_i=k} (\mathbf{y}_i - \tilde{\mu_k})(\mathbf{y}_i - \tilde{\mu_k})^T$$

$$\tilde{\mathbf{S}}_{\mathbf{B}} = \sum_{k=1}^{K} N_k(\tilde{\mu_k} - \tilde{\mu})(\tilde{\mu_k} - \tilde{\mu})^T, \qquad (2.5)$$

$$where \quad \tilde{\mu_k} = \frac{1}{N_k} \sum_{c_i=k} y.$$

The within class scatter matrix is the covariance for the difference between each data point and the mean vector for its corresponding cluster across all clusters. The between class scatter matrix finds the covariance for the difference between each cluster's mean vector and the overall mean for all data points.

While PCA maximizes global variation of the data, LDA maximizes the variation specific to the clustered architecture since class labels are known. Intuitively, by separating the data into clusters and maximizing the ratio of between to within cluster variation, the resulting projection $\mathbf{W}^*_{\text{LDA}}$ will be one which maximizes the variance between clusters while minimizing the variance within each cluster. LDA assumes that all classes have equal covariance, but the extension to quadratic discriminant analysis (QDA) allows for unequal covariance matrices when calculating the within class scatter matrices, therefore allowing for more flexible, non-linear decision boundaries between classes.

LDA is a supervised classification technique which requires class labels in order to maximize this ratio of between class variation to within class variation. The challenge of bridging the gap between PCA and LDA is defining these class labels which capture latent discrete structure of data. The core idea of iDA is to use an

unsupervised clustering method to define classes and use those in LDA to find a low dimensional embedding. This unsupervised method aims to find a latent clustered structure in the data while being unbiased to factors such as batch, population, or cell type. This agnostic approach avoids using recorded experimental class labels, but the iDA algorithm does allow for an initial set of class labels to be given if desired by the user. Iteration in this case would start with the embedding step instead of clustering step using the given labels, but that will just find an embedding that separates based on the provided labels. An approach that incorporates labels such as soft-assignments would be helpful, but we leave it for future work.

2.2 iDA Algorithm

The iDA algorithm determines the number of clusters to use in each iteration by maximizing the clustering's modularity as used in the Walktrap clustering step. If, however, the user would like to find an embedding with a predetermined number of clusters, this can be achieved by setting the user-defined $c.param$ parameter as input to the iDA function.

For batch correction, instead of only providing known batch indicators as input class labels to LDA, we assume variation may come from unmeasured sources, and therefore using a variable like batch would not fully capture groups arising from such variation. Moreover, in the case of scRNA-seq, cell types are unknown and since known marker genes are not defined for all cell types as well as subtypes, inferring cell types is difficult.

Algorithm 1: iDA algorithm

Input: Data matrix $\mathbf{X}$
Output: Embedding $\mathbf{Y}$
Initialize: choose set M of informative features to initialize embedding $\mathbf{Y}_0 = \mathbf{X}_{\cdot,M}$;
for $t = 1, 2, \ldots$ do
  if *converged* then
    return $\mathbf{Y}_{t-1}$
  end
  Use Network-based clustering of $\mathbf{Y}_{t-1}$ to get class assignments c_i that maximize cluster modularity;
  Obtain $\mathbf{W}^*_{\text{LDA}}$ by maximizing $J_{\text{LDA}}(\mathbf{W}; \mathbf{Y}_{t-1})$;
  Set $\mathbf{Y}_t = \mathbf{Y}_{t-1}\mathbf{W}^*_{\text{LDA}}$
end

The iDA algorithm is as follows: 1) Extract informative genes to identify those which may be potentially affected by unmeasured variables, 2) Cluster the samples using the Louvain community detection method to get class labels for each sample, 3) Decompose the informative gene expression matrix via Linear Discriminant Analysis (LDA) or Quadratic Discriminant Analysis (QDA) using the class labels obtained from the Louvain clustering. 4) Update the class labels by clustering in the discriminant space. The new optimal class assignments can then be used for decomposition until the class assignments have stabilized in the embedding space.

iDA results in an embedding space of dimension $K - 1$, where K is the number of clusters identified by iDA which maximizes the community structure modularity, and each dimension determines a decision boundary which best separates an individual class from the rest. To find reliable clusters which define the discrete latent structure, we use the Walktrap method for community detection in a nearest neighbor graph which computes the similarities between neighbors based on random walks in the graph (Pons and Latapy, 2005). To compute the shared nearest neighbors graph, first,

a k-nearest neighbors graph is built with the default number of neighbors $k = 10$. Then, the pairwise Jaccard Similarity is computed (the number of common neighbors each vertex has divided by the number of vertices that are neighbors of at least one of the two pairwise vertices). The shared nearest neighbors graph is then pruned such that pairwise vertices that have similarities less than 1/15 are set to zero, and converted to an undirected graph, which is then the input to the Walktrap clustering algorithm. The Walktrap clustering returns a hierarchical clustering. The point at which to cut the tree is determined by computing the modularity at each level of the tree and cutting it where the modularity is maximized.

For our experiments below, we check for convergence as explained in Section Convergence Assessment. Since iDA remains agnostic to the study design, we don't need to worry about recording errors for phenotype or batch data, or self-identification discrepancies. In our implementation we initialize the embedding by selecting M features with highest marginal variance.

2.3 Methods for Performance Assessment

To assess the performance of iDA, we used two data sets. The first is from the Geuvadis bulk RNA-sequencing Project, which has expression data from 462 individuals from European and Yoruba ancestry individuals and was sequenced in 7 different laboratories (Brown et al., 2018). This data set exhibits a latent structure from an overwhelming amount of variation attributed to batch effects and some

variation attributed to population structure. The second data set is the 10X Genomics peripheral blood mononuclear cell (PBMC) single cell RNA-sequencing (scRNA-seq) data set which has expression from roughly 3,000 unlabeled immune cells (Zheng et al., 2017). We chose these datasets because of the known structure embedded in each (genetic population structure and batch in the Geuvadis and distinct cell types in PBMC dataset). Note, however, these labels are not used as input to the iDA algorithm.

2.3.1 Convergence Assessment

To assess the convergence of iDA, we need to inspect the stability of the cluster assignments resulting from the Walktrap community detection at each iteration. Since the cluster assignments dictate the scatter matrices calculated for the discriminant analysis, once the cluster assignments are stable, then the discriminators will not change. To evaluate when the cluster assignments have stopped changing, we calculate the Adjusted Rand Index as concordance for the cluster assignments between each iteration. The iterations are stopped when the Adjusted Rand Index is larger than 0.98 between the cluster assignment vectors.

2.3.2 Cumulative F-Statistic

To compare the capacity of iDA and PCA to accurately find a subspace which describes the underlying clustering in the data, we adapt the F statistic as well as R^2 to be able to assess models across dimensions. First, we perform linear regressions

which model each dimension of PCA and iDA as a function of the observed variables that describe the known underlying group structure (for Geuvadis data set, this is the interaction between population and laboratory, for 10X Genomics PBMC scRNA-seq data, this will be cell type). To determine the appropriate number of PCs to use in this comparison for each dataset, we evaluated the elbow plots to determine the eigenvectors accounting for the majority of the variation in each dataset (Figure 2.2).

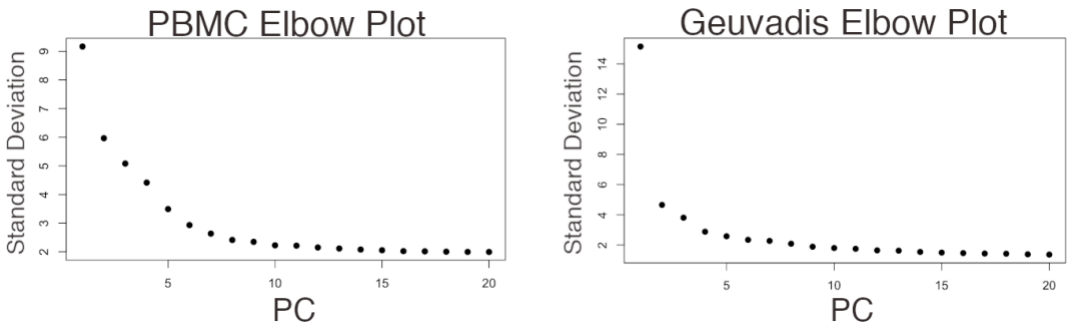


Figure 2.2. Elbow plots to determine the optimal number of dimensions to use in the PCA embedding.
To determine how many top PCs to use for each dataset, the eigenvalues are plotted for each PC dimension, and the cutoff is determined based on the "bend" in the plot. To ensure an appropriate amount of variation is accounted for when comparing to iDA, we chose to include the top 10 PCs in the Geuvadis and PBMC datasets.

To ensure an appropriate amount of variation was captured in the PC embedding for comparison to iDA, we use the first 10 PCs for the Geuvadis and PBMC datasets.

$$\underbrace{X_i}_{\text{dimension i of Geuvadis reduction}} = \underbrace{B_0 + B_1 population : laboratory}_{\text{underlying group structure}} \tag{2.6}$$

$$\underbrace{Y_i}_{\text{dimension i of scRNA-seq reduction}} = \underbrace{B_0 + B_1 celltype}_{\text{underlying group structure}}$$

Then, to assess how well these known classes are separated across dimensions, we use an adaptation of the F-statistic. We define this cumulative F-statistic as:

$$cumulativeFstat = \frac{\sum_{i=1}^{k} TSS_i - \sum_{i=1}^{k} RSS_i}{\sum_{i=1}^{k} RSS_i} \tag{2.7}$$

where we compute the total sum of squares (TSS) and residual sum of squares (RSS) per regression model, and sum over each dimension to compute a cumulative F-statistic from 1 to k. This statistic measures the variance of the group means / variance of the within group variances over dimensions. This indicates how well the dimensional reductions for PCA and iDA describes group separation for known groups. Because both the $\mathbf{TSS}$ and $\mathbf{RSS}$ change as dimensions are added, the cumulative F-statistic is not monotonic.

2.3.3 Cumulative R^2

We also compute a cumulative R^2 for dimensions 1 through k to measure the total variance captured in the sub-spaces from each reduction technique over dimensions.

$$cumulativeR^2 = 1 - \frac{\sum_{i=1}^{k} RSS_i}{\sum_{i=1}^{k} TSS_i} \tag{2.8}$$

This cumulative R^2 over dimensions represents the proportion of variance over dimensions that is explained by the underlying grouping.

2.3.4 Clustering Validation

To validate that the iDA method identifies better latent clustering in the original data than the PCA dimensionality reduction and is not creating spurious clustering, we will assess the compactness and separation of clusters in the original space using the Dunn Index (DI). The DI is defined as the ratio between the smallest distance between observations not in the same cluster to the largest intra-cluster distance,

$$\min_{1\leq i\leq c}\left\{\min_{1\leq j\leq c, j\neq i}\left\{\frac{\delta(X_i,X_j)}{\max\limits_{1\leq k\leq c}\Delta(X_k)}\right\}\right\} \tag{2.9}$$

where $\delta(X_i, X_j)$ is the inter-cluster distance and $\Delta(X_k)$ is the intra-cluster distance.

To ensure robustness, the Dunn Index is calculated using a bootstrapping approach over multiple sampled observations samples drawn with replacement and the Dunn Index is calculated over 1000 bootstrap samples.

2.4 Results

2.4.1 iDA preserves cluster structure better than PCA

The cumulative F-statistic is a measure to evaluate how well the low dimensional embedding captures the separation of clusters in the data. The iDA algorithm was applied to the 10X genomics scRNA-seq PBMC data set and compared to the PCA

reduction of the same dimension. For comparison of the embedding from iDA and PCA, we used the cell identities assigned from the Seurat scRNA-seq pipeline as the cell types.

Figure 2.3A provides evidence that the decision boundaries from iDA offer an embedding which better separates the putative cell types compared to PCA, as determined by the cumulative F-statistic. As expected, the first dimension of PCA, which captures the most variation in the data, separates one cluster very well (likely the most variable cluster), but as dimensions are added, the projections do not capture the cluster separation. In the iDA embedding, we see each dimension equally captures the separation of one cluster, so the cumulative F-statistic does not degrade as we add dimensions. iDA also finds LD's which capture a higher proportion of variance explained over all dimensions than the PCs do for evaluating variance between cell types (Figure 2.3B). Although PCA's objective function is to maximize variance, we see that it fails to find variation attributed to some cell types, likely ones with low variation. There is evidence of this in the acute drop off in both the cumulative F-statistic and R^2 values.

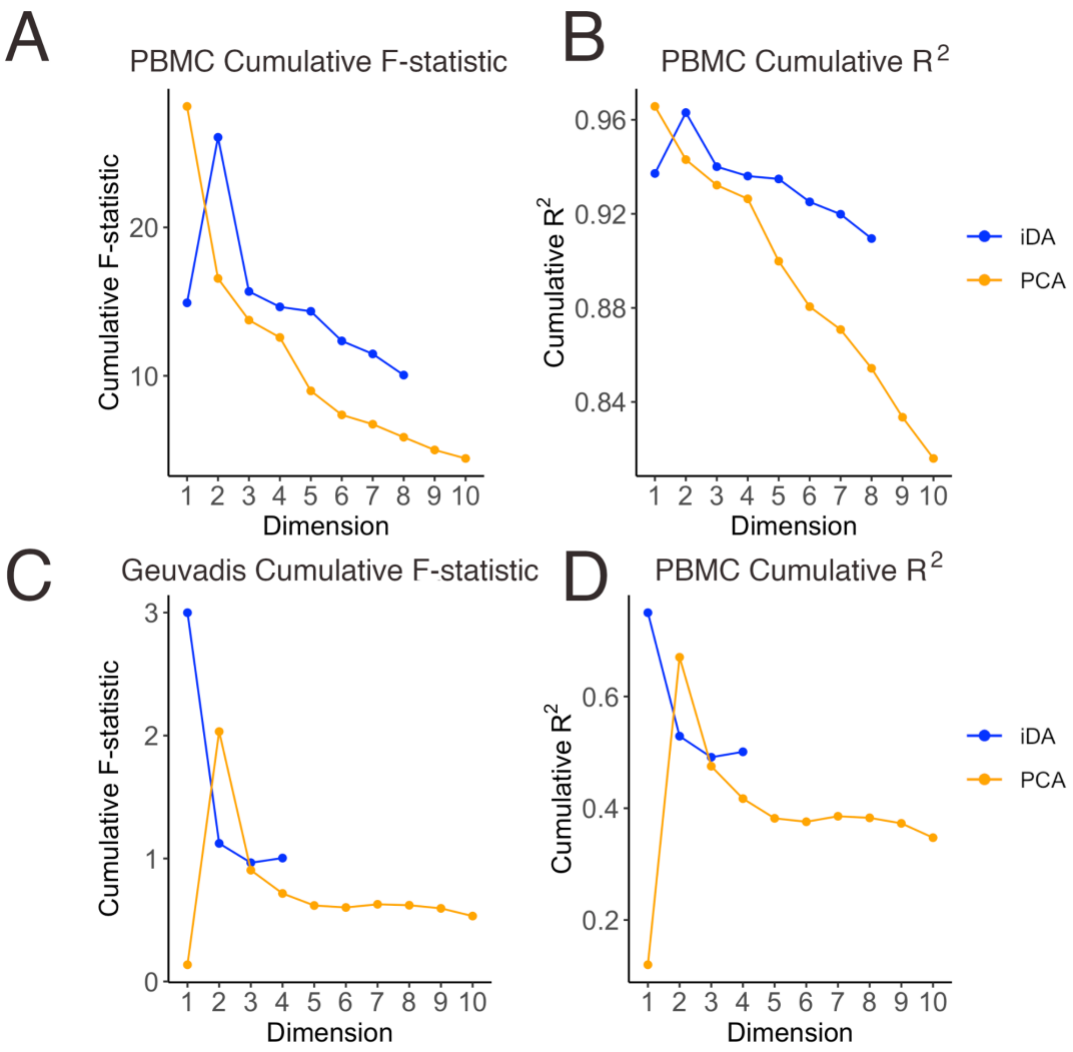


Figure 2.3. Cumulative F-statistic and Cumulative R^2 for the PBMC and Geuvadis datasets.

A) As more dimensions are added for each of the reduction methods, the clustering from as determined by the Seurat putative cell type assignments for each of the samples is better separated by the iDA embedding than PCA. B) The iDA embedding captures more of the total variance attributed to the latent clustering than PCA. C) As more dimensions are added for each of the reduction methods, the clustering from the main effects and interaction between laboratory and population for each of the samples is better separated by the iDA embedding than PCA. D) The iDA embedding captures more of the total variance attributed to the latent clustering than PCA.

We see a similar trend for the statistics computed using the Geuvadis data set. We evaluate the performance of iDA and PCA using the groups from the main effects and interaction between the population and laboratory the samples were processed in. While these variables may not account for all latent clustering in this data, we know they contribute to a large portion of the clustering, and an embedding should capture these clusters (and potentially others). Here, the first dimension of PCA does a very poor job of accounting for the latent clustering attributed to population and laboratory; more evidence that maximizing the variation does not guarantee to best separate clusters (Figure 2.3C). iDA again consistently finds better separation for this data, as well as better overall variation attributed to these groups (Figure 2.3D).

Many of the clusters produced by both iDA and PCA + Louvain (Seurat clusters) are nearly identical between each method (Table 2.1). The main difference is the boundary between Seurat Clusters 0 and 1 changes such that some Seurat Cluster 0 cells get grouped with cells in Seurat Cluster 1.

Table 2.1 Confusion matrix between iDA and PCA+Louvain Clustering

iDA Cluster	PCA + Louvain Cluster									
	0	1	2	3	4	5	6	7	8	9
1	0	0	426	0	0	2	0	0	0	2
2	3	1	0	0	176	0	3	142	0	0
3	523	2	0	0	14	0	0	0	0	0
4	120	454	0	1	1	0	0	0	1	0
5	0	0	0	343	0	0	0	1	0	0
6	0	0	0	0	4	0	152	10	0	0
7	0	0	18	0	0	192	0	0	0	0
8	0	0	1	0	0	0	0	0	33	0
9	0	0	0	0	0	0	0	0	0	13

Based on both the tSNE and UMAP of these projections, the boundary between these two clusters is not well defined (Figure 2.4). To determine which clustering is optimal, we compute the bootstrapped Dunn Index (DI) in the full gene expression space for class assignments resulting from iDA compared to the class assignments from other methods over 1,000 iterations. The PCA and iDA embeddings were compared using multiple clustering techniques including two community detection algorithms common in single cell clustering, Walktrap clustering (the default for the iDA algorithm), and Louvain Community Detection (Blondel et al., 2008), as well as K-means clustering, which has been shown to be closely linked to the unsupervised dimension reduction of PCA (Ding & He, 2004).

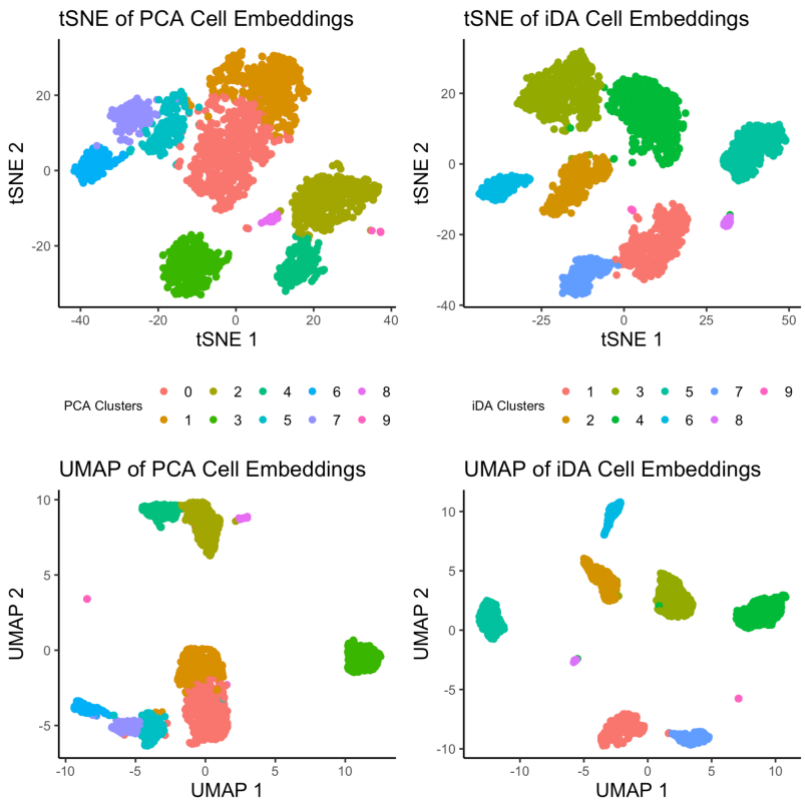


Figure 2.4. tSNE and UMAP visualizations for the PCA and iDA PBMC cell embeddings.
Similarly to PCA, the iDA embedding can be used with both tSNE and UMAP to visualize cluster separation.

The clusters found in the PBMC scRNA-seq dataset using the iDA embedding with each of the community detection algorithms -- Walktrap (green) and Louvain clustering (light blue) -- both have Dunn Indices which are significantly increased compared to the clusters found with the PCA embedding and Louvain clustering (i.e. the clusters found with the Seurat pipeline) (yellow) (Figure 2.5A). In the Geuvadis dataset, the Dunn Indices of the clusters found by all clustering methods with the iDA embedding are all significantly increased compared to the clusters found by both the PCA embedding with Louvain clustering as well as the recorded batch (laboratory)

and population of each sample (Figure 2.5B). This indicates that the class assignments resulting from iDA define clusters in the original space which are compact and better separated than class assignments determined by either the Seurat pipeline for scRNA-seq or measured covariates in Geuvadis.

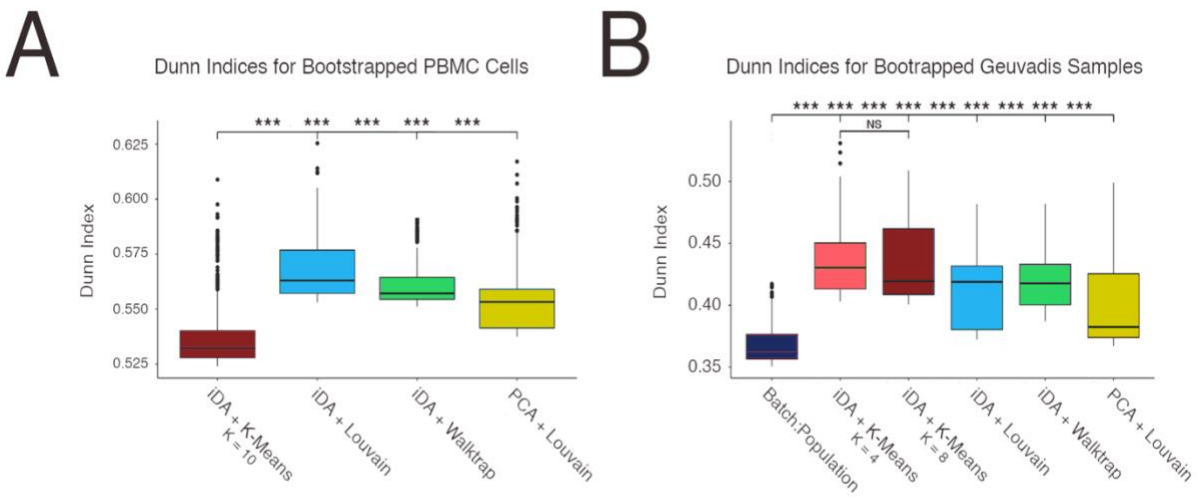


Figure 2.5. Dunn Index calculated for bootstrapped samples for each dataset. Dunn Index computed for bootstrapped samples from PA) PBMC cells and B) Geuvadis samples using either the PCA embedding or the iDA embedding and several different clustering methods. The sampling was performed by taking n = number of observations in each dataset, with replacement, and bootstrapped 1,000 times. A) The iDA algorithm with the Community detection algorithms (Walktrap and Louvain) both have significantly higher Dunn Indices than the PCA embedding with Louvain clustering. B) All clustering methods with the iDA algorithm generate significantly better clusters with respect to the Dunn Indices when compared to the PCA embedding and the known batch and population.

2.4.2 iDA directions are highly correlated with known PBMC cell type markers

Canonical markers of particular cell types have been characterized in PBMCs and can be used to validate clustering methods (Zheng et al., 2017). These markers have been experimentally elucidated such that cells with high expression of these markers are likely of the corresponding cell type based on their known function.

CD14 and LYZ are known markers for monocytes, MS4A1 is a marker for B-cells, GNLY and NKG7 are markers for Natural Killer (NK) cells, and PPBP is a marker for platelets. If the clustering algorithm captures clusters indicative of cell type, the resulting clusters should reflect this in the marker gene's expression. For example, LD6 is the directional vector which separates cluster 1 (with cells in cluster 1 mapping to the lowest weights of the vector). We see that the cells in cluster 1 highly express the two markers for Monocytes, CD14 and LYZ (Figure 2.6A). Quantitatively, both markers for monocytes have much higher mean scaled expression than the cells in all other clusters (Table 2.2).

Table 2.2. Comparison of mean expression of canonical PBMC cell type markers.

Difference in Mean Expression for Canonical PBMC Cell Type Markers

Cell Type	Marker Scaled Expression	Mean Expression Within Cell type	Mean Expression All Other Cells
Monocytes	LYZ	1.668	-0.382
Monocytes	CD14	1.431	-0.328
NK Cells	GNLY	2.740	-0.150
NK Cells	NKG7	2.436	-0.133
B-Cells	MS4A1	1.971	-0.291
Platelets	PPBP	9.792	-0.053

The LD3 vector separates the cluster of cells (cluster 6) which highly express the markers of NK cells (GNLY and NKG7) (Figure 2.6B). These markers are shown to have much higher average scaled expression than cells not in this cluster. Additionally, iDA is sensitive to separating the cluster of cells which are high in the expression of both marker genes (as opposed to only one), separating them from cells in cluster 2 which are high in only NKG7 expression but not GNLY, indicating that cluster 6 is a cluster of NK cells.

LD5 separates the cluster which highly expresses the marker for B-cells, MS4A1 (Figure 2.6C) , also with an overall much higher mean scaled expression when compared to cells in all other clusters (1.971 versus -0.291). This indicates cluster 5 is likely a population of B-cells.

Lastly, a marker for platelets, PPBP, is highly expressed in the cells in cluster 9 identified by the iDA algorithm. This cluster is separated by LD1 and shows a clear separation, with the platelet cells mapping to high LD weights (Figure 2.6D).

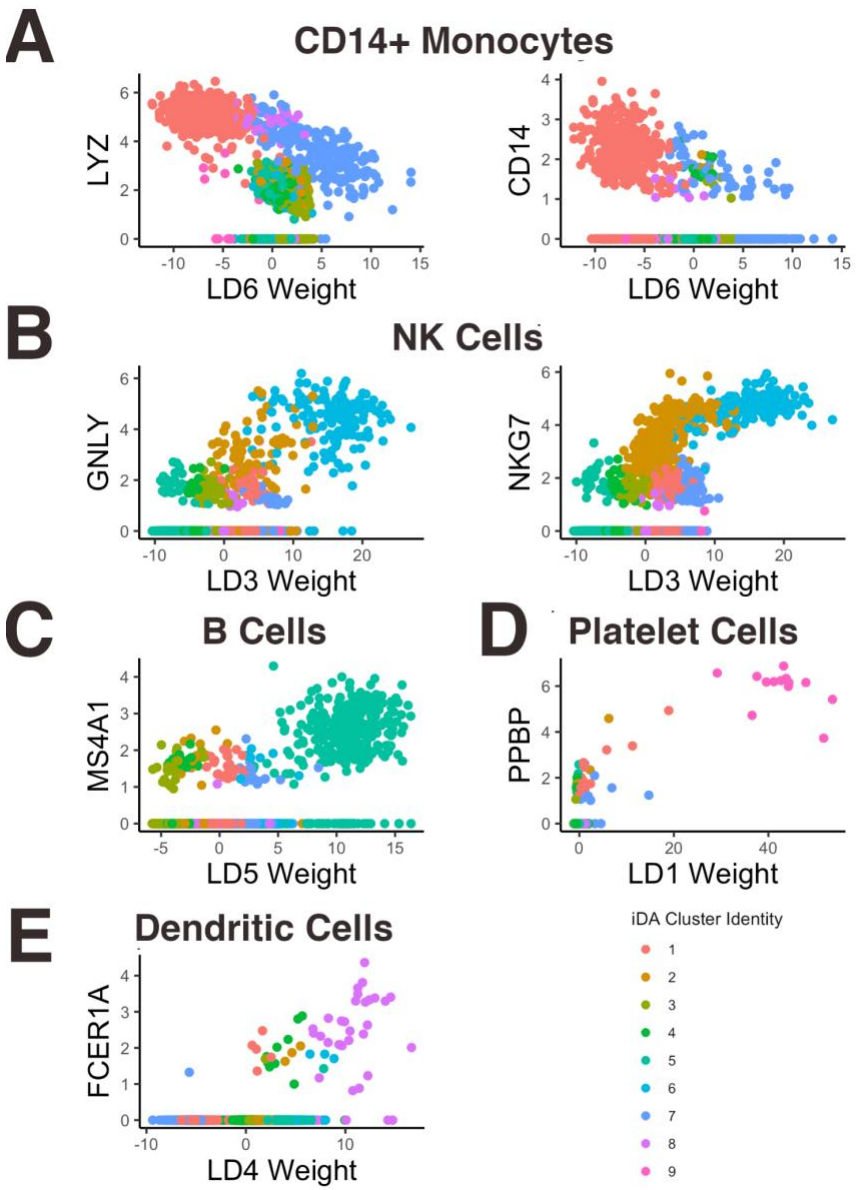


Figure 2.6. Putative cell type clusters found with iDA are separated both by known cell type markers and in the iDA embedding.
A) The group of cells which highly express both LYZ and CD14, known markers for monocytes, are separated by the $6th$ iDA dimension (LD6) (salmon colored points). B) The group of cells which highly express GNLY and NKG7, known markers for NK cells, are separated by the $3rd$ iDA dimension (LD3) (light blue colored points). C) The group of cells which highly express MS4A1, a known marker for B-cells, are separated by the $5th$ iDA dimension (LD5) (teal colored points). D) PPBP is a known marker for platelets. The group of cells which highly express PPBP are separated by the $1st$ iDA dimension (LD1) (pink colored points). E) FCER1A is a known marker for Dendritic cells. The group of cells which highly express FCER1A are separated by the $4th$ iDA dimension (LD4) (purple colored points).

2.4.3 iDA directions are highly correlated with SNPs in eQTL

The Geuvadis data set has known sub-population features attributed to ancestry and the location where the samples were sequenced. iDA identifies an LD which separates the YRI population from the EUR populations (LD3). Interestingly, this LD purely captures the separation in population and groups all samples sequenced in different laboratories together.

Brown et al (2018) show that expression reflects population structure because alleles, which may have different frequencies among different populations, can affect the expression level of nearby genes. These allele-gene pairs form expression quantitative trait loci, or eQTL, if this relationship occurs. The Geuvadis data set has 116 SNP-gene pairs which are in eQTL only in YRI individuals and not in the EUR individuals. The average difference between the YRI minor allele frequency (MAF) and the EUR population's MAF is a substantial 0.175. Since these alleles are in eQTL with a gene pair and the allele frequencies are different between populations, the effect on expression level will also differ between populations. These SNPs allele frequencies vary greatly between the YRI and EUR populations (Table 2.3).

Table 2.3. Minor Allele Frequencies per population of top eQTL SNP/gene pairs.

SNP	Gene	AFR MAF	EUR MAF
rs11757158	RAB5C	0.45	0.09
rs114313536	SRF	0.16	0.02
rs143415501	PSKH1	0.06	0.00
rs200846953	PWAR6	0.54	0.83

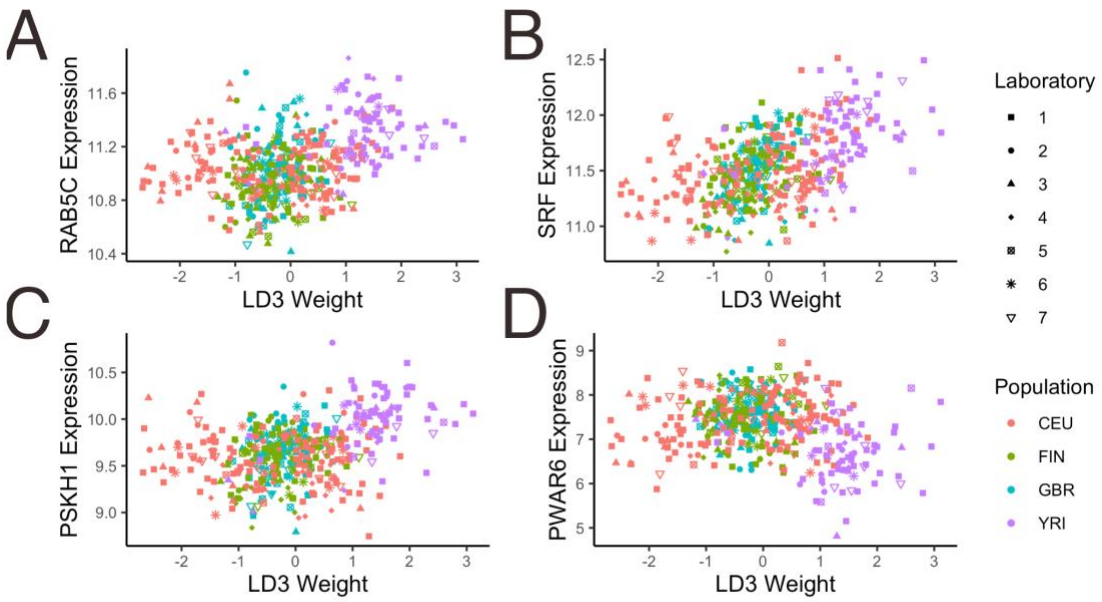


Figure 2.7. iDA projection 3 separates populations.
SNP-gene pairs in eQTL affect the expression dependent on the sample population since the SNP MAF is variable across populations. The $3rd$ iDA dimension (LD3) separates the YRI population from the European ones. A) RAB5C, which is in eQTL with SNP rs11757158 in the YRI population, has higher average expression than the European populations. B) SRF, which is in eQTL with SNP rs11413536 in the YRI population, has higher average expression than the European populations. C) PSKH1, which is in eQTL with SNP rs143415501 in the YRI population, has higher average expression than the European populations. D) PWAR6, which is in eQTL with SNP rs200846953 in the YRI population, has higher average expression than the European populations.

The MAFs differ as much as 36% (as seen in the rs11757158-RAB5C pair), and in some cases, the minor allele is not present at all in the EUR population (rs143415501). Figure 2.7 A, B and C clearly increase the expression rate of their respective eQTL genes, while D shows a decrease in expression of PWAR6 in YRI as compared to the EUR samples, mimicking the difference of MAF between the YRI population and the EUR populations.

2.4.4 iDA directions are highly correlated with known technical noise

Another point to consider is if technical noise is being further entrenched in the embeddings found by iDA. To ensure iDA is able to capture variation in datasets with unknown technical noise, we performed iDA on a control dataset designed with all variation attributed to technical noise (Zheng et al., 2017). This dataset consists of droplets which were loaded with the same ratio of 92 External RNA Controls Consortium (ERCC) synthetic RNAs into GEMs. Since there are no differences in cell size, RNA content or transcriptional activity in these GEMs, the only variation is from technical noise introduced by the differing number of UMIs per GEM. The embedding iDA produces on both the raw counts and the normalized counts confirms that the entrenchment of technical noise can arise from the normalization technique (Townes et al., 2019). For example, in the negative control ERCC dataset, the iDA embedding of the raw counts clearly separates GEMs by the technical variation introduced by the mean UMI count per GEM (Figure 2.8B), but the embedding of the log normalized counts does not produce this separation (Figure 2.8A). Therefore, the entrenchment of technical noise in the iDA embedding is a result of the normalization step, and not the iDA algorithm itself.

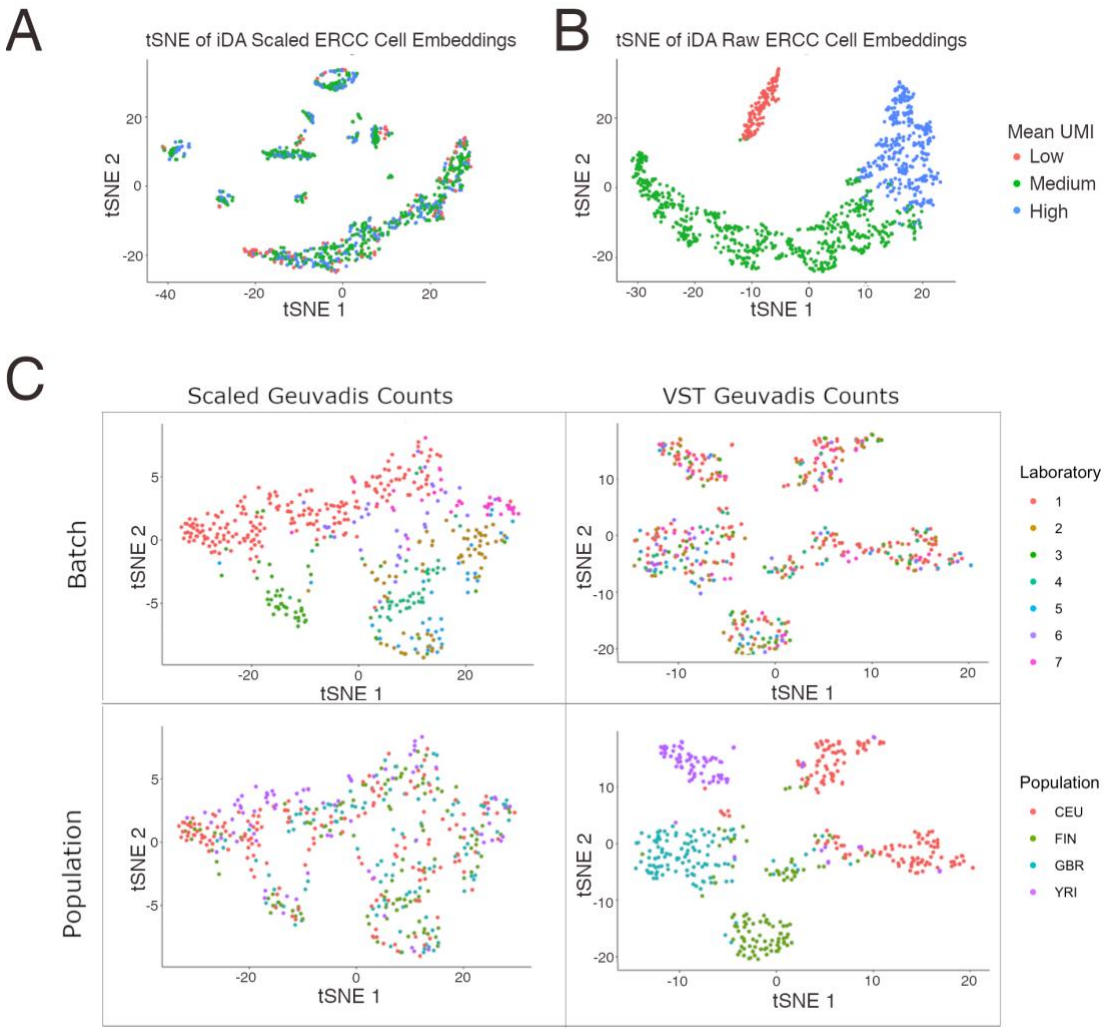


Figure 2.8. tSNE of iDA embedding for ERCC negative control dataset.
tSNE of the iDA embedding for the ERCC negative control dataset both on the scaled counts (A) and raw counts (B). iDA captures the variation attributed to the technical noise when present in the data (B). If normalization techniques have entrenched the technical noise (Townes et al., 2019), iDA will not be able to separate the technical variation (A).

To show further evidence of accurate iDA clustering and dimensionality reduction, we applied both iDA and PCA with Walktrap clustering to two single cell benchmarking datasets from the 10X genomics platform with known cell type labels (Tian et al., 2019). Known populations of three (10X-3cl) and five cell lines (10X-

5cl), respectively, were sequenced using the 10X genomics single cell RNA sequencing platform for these two datasets. The tSNE plot of the PCA reduction of log counts for each cell reveals discrete clustering within each cell line, but there is also a moderate amount of heterogeneity in each cell line (particularly in H1975 and A549, which each separate into distinct sub-clusters) (Figure 2.9A). When iDA is applied to each of these datasets without any guidance on how many cell lines are in each dataset, the resulting clustering is sensitive to the cell type sub-clustering. Because of this sensitivity to within cell type heterogeneity, when compared to the known cell type labels, the Adjusted Rand Index (ARI) for iDA clustering on the 3 cell line dataset is .519, and .798 for the 5 cell line dataset. Comparatively, when unsupervised clustering is applied to the PCA embedding, the ARI between these clusters and the known cell type labels is .457 for 10X-3cl and .615 for 10X-5cl.

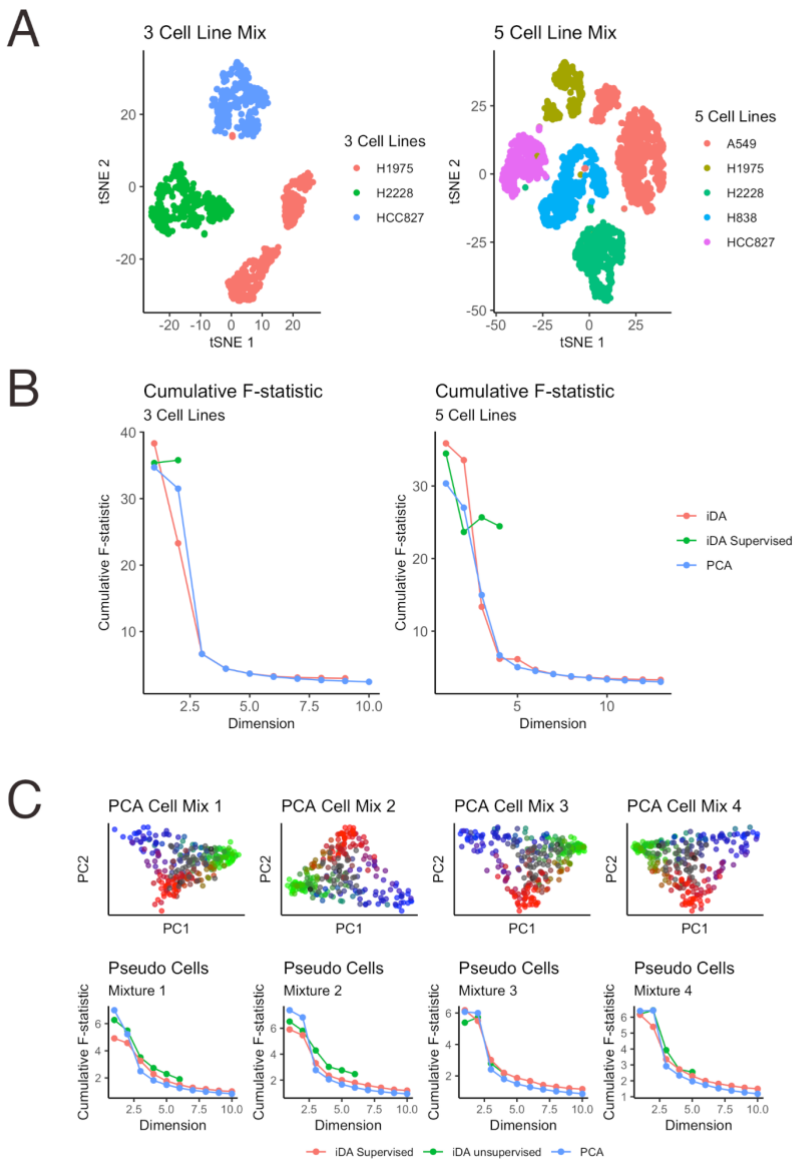


Figure 2.9. Method comparison on pseudo-cell scRNA-seq dataset.
A) tSNE of the Tian et al. (2019) 10X genomics single cell RNA sequencing benchmarking datasets. Distinct clustering is visualized for each of the known cell lines, but there is a substantial amount of heterogeneity within each cell line for both datasets. B) Cumulative F-statistics for the iDA supervised embedding, iDA unsupervised embedding, and PCA embedding. The iDA supervised embedding outperforms the other methods in capturing variance attributed to the known cell line labels. C) Pseudocell mixtures with different ratios of RNAs from 3 cell lines. Visually, PCA separates the homogeneous pseudo cells (red, green, blue), but the mixed populations form a continuous spectrum between the 3 homogeneous ones. The cumulative F-statistic for each dataset shows iDA unsupervised, iDA supervised, and PCA all perform comparably in these in-discrete datasets.

In the case where the number of cell lines in a sample is known, this parameter can be given to Walktrap clustering. In these two datasets, because there is such a great difference between the expression profiles of each cell line, the supervised clustering yields much greater results. When iDA is applied with the number of known clusters given, the resulting ARIs are .997, .986 for 10X-3cl and 10X-5cl respectively. PCA with supervised clustering also does well with ARIs of 1.00 for 10X-3cl and .987 for 10X-5cl (Table 2.4).

Table 2.4. Adjusted rand index for clusters from iDA versus PCA embeddings.

Dataset	iDA Supervised ARI	PCA Supervised ARI	iDA Unsupervised ARI	PCA Unsupervised ARI
10X-3cl	.996	1.00	0.519	0.457
10X-5cl	0.986	0.987	0.798	0.615

2.4.5 iDA Converges

When assessing the convergence of iDA, we evaluate the concordance of cluster assignments at each iteration with the Adjusted Rand Index between the two sets of class assignments. We set a threshold of concordance larger than 0.98 as the stopping criterion for iDA. Each data set reaches the threshold within 3 iterations of iDA (Figure 2.10). Since the discriminants in iDA are determined based on the cluster assignments, if the assignments reach stability between iterations, the discriminants will be stable and unchanged between iterations as well.

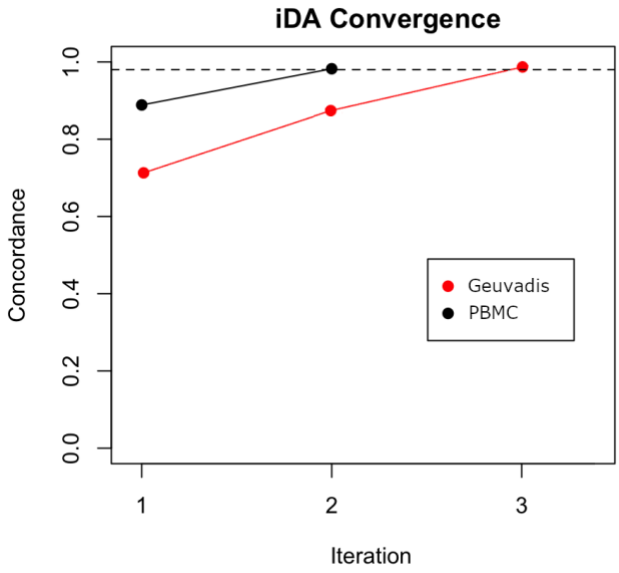


Figure 2.10. iDA Convergence.
Convergence of the iDA algorithm is measured by the concordance of class assignments of each iteration to the previous one. The convergence threshold for iDA is set to concordance = 0.98. Both the Geuvadis and PBMC data sets converge within 3 iterations.

The speed of iDA is heavily dependent on the number of variable features used initially since each iteration decomposes the gene-by-gene matrix. Although PCA is faster, iDA runs within 6 minutes for datasets with less than 3,000 variable genes and 500 samples. iDA identifies 2,829 and 3,000 variable genes in the Geuvadis and PBMC datasets and runs in five and three minutes, respectively (Figure 2.11). In the breakdown of time allocation for the iDA algorithm, the SVD takes the most time for each of the data sets, since a decomposition is computed in each iteration.

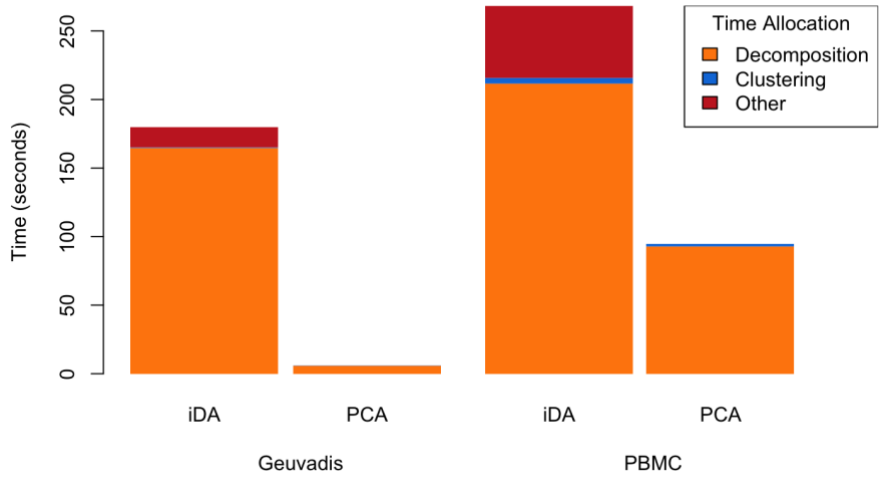


Figure 2.11. iDA runtime.
Run time of the iDA algorithm versus the PCA reduction for both the Geuvadis and PBMC data sets. PCA is much faster than iDA, as it only computes one decomposition. iDA computes a decomposition and clustering in each iteration until convergence. The Geuvadis data set which has 462 samples and iDA identifies 1,802 variable genes runs within 3 minutes. The PBMC data set has 2,638 cells and iDA identifies 3,000 variable genes and runs within 4.47 minutes.

2.5 Discussion

Discovering the latent structure in high dimensional data is essential in many applications. The latent structure can be attributed to batch, which needs to be corrected for before downstream analysis to avoid spurious associations as well as reduced power in association testing. The latent structure can also define biological clusters which researchers are interested in evaluating such as in scRNA-seq, where identifying a set of features which define cell types is of main concern. In standard pre-processing using PCA, as well as methods which use PCA to identify sources of variation in batch effects cases, classes which are distinctly defined but have low variation will not be maximally separated. Although both PCA and iDA are not specifically batch correction methods, the resulting low dimensional embedding

which captures unwanted variation can in turn be used as covariates in downstream analysis to produce more accurate results. For example, the iDA embedding which captures the separation of batches in the scaled Geuvadis counts clearly separates the samples from individual laboratories in which they were processed and could be used as covariates in downstream analysis to account for this separation (Figure 2.7C). This approach, namely, adding embeddings to control for unwanted structure in downstream analysis, is preferred to another common approach of removing the first few embedding directions from the original input data to produce corrected datasets for further downstream analysis. Accordingly, we have focused our analyses in this paper on the preferred approach.

One advantage of PCA embeddings is the interpretability of the variability contribution each dimension adds when included in the embedding. Often, the respective eigenvalues of the top PCs are plotted to determine the "elbow" which determines the ideal number of top PCs to include in the final embedding (Figure 2.2). This offers a natural ordering of PCs based on the amount of variation the PC accounts for in the data. Because iDA also finds embeddings with an SVD, the resulting LD's order corresponds to “easy-to-separate” clusters. For example, in the PBMC dataset, LD1 separates the platelet cells which are best separated from other cell types (Figure 2.5D), while the later LDs are not as well separated (Figure 2.5 A,C,E).

Several neighbor-preserving non-linear dimensionality reduction techniques exist for data visualization such as t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). These methods are particularly popular for scRNA-seq dimensionality reduction and visualization. tSNE constructs a probability distribution for all pairs of samples in the data such that samples close in Euclidean distance in the original data space to have a higher probability of being placed near each other in the resulting lower dimensional space than those which are far apart from each other in the original data space, therefore preserving the neighborhood structure in a two or three dimensional space (Laurens van der Maaten & Geoffrey Hinton, 2008). tSNE has a few shortcomings, particularly when applied to scRNA-seq. tSNE does not scale well which is problematic with the rapidly growing number of cells being sequenced for scRNA-seq. iDA is more scalable for large numbers of samples since the SVD time-consuming step in iDA is dependent on the number of variable features and not the number of samples. Since tSNE computes the k-nearest neighbor graph as an initial step, it is infeasible to compute this in the original high dimensional space, so a preprocessing step like PCA is used to reduce the number of dimensions before applying tSNE, which means it could fall victim to the downfalls of PCA as discussed before. The newer method, UMAP, is also a neighborhood-preserving method with several adaptations to tSNE, such as a cost function which attempts to preserve global structure (McInnes et al., 2018).

While t-SNE or UMAP may suffice for data visualization, the embedding only offers visual information on dissimilarity between the samples and loses the interpretability that the iDA embedding offers. The tSNE and UMAP embeddings do not have weights of the features which define the differences between samples, which renders the embeddings useless to explain features which drive cluster separation or define features that may be used as covariates in downstream analysis. From the SVD of $\mathbf{S_W^{-1}} * \mathbf{S_B}$ yields the weight matrix of genes by linear discriminants. These gene weights allow for the interpretation of specific features which separate latent clusters.

With these interpretable embeddings, we can visualize the separation of the clusters in each dimension of the embedding and validate these clusters with known properties of each group. For instance, to validate the putative cell types found in the scRNA-seq data set, we compare the expression of known markers to LDs which separate the identified cell type clusters. The cell type clusters identified by the iDA algorithm show corresponding levels of expression for known markers of these cell types. The iDA embedding and clustering is also sensitive to lowly-abundant cell types. In PBMCs, lymphocytes (T cells, B cells, NK cells) are the majority of the cell population, monocytes are the next highest abundant, and dendritic cells and platelets as the rarest (Kleiveland, 2015). In the iDA embedding of the PBMC dataset, the least abundant platelet cell population is the cell cluster which is best separated from the others (in LD1), and the dendritic cell population is captured in LD4 (Figure 2.5). Therefore, iDA will identify rare populations that are sufficiently separated.

Expression is known to reflect population structure because of the eQTL (Brown et al., 2018) so it is essential to identify this source of variation as a covariate if this is not the desired response variable. In the Geuvadis data set, of the 116 SNPs which are in eQTL in only the YRI individuals, the average difference between the YRI MAF and the EUR population's MAF is a substantial 0.175. This means that on average, SNPs in eQTL in the YRI population but not in the EUR populations have a frequency difference in the population of 17.5%. Since these SNPs are in eQTL in only the YRI population, this means they are affecting the expression of the gene which is in eQTL with them much differently than other populations. The difference in population structure affecting MAFs means that the expression of genes in eQTL will be affected by the population structure as well, as seen in Figure 2.7, causing a latent clustering of expression. The genes in eQTL with population-dependent alleles are recapitulated in the separation of populations by LD as well as gene expression.

When comparing the embeddings that iDA produces compared to PCA, the cumulative F-statistic shows the iDA supervised embedding outperforms the other methods in both datasets with respect to the known cell types. The iDA unsupervised embedding outperforms its PCA counterpart in 10X-5cl and does just as well as the PCA embedding in 10X-3cl (Figure 2.9B).

In instances where the nature of the latent structure is from discrete sources of variations such as batch effect, population stratification, and discrete cell types, iDA is able to capture these clusters well in the resulting embedding and clustering. iDA

leverages LDA in its algorithm to find the embedding and therefore is not as effective at finding an accurate embedding in scenarios where the latent structure is of a continuous fashion, such as in developmental biology. In such datasets where cell types are differentiating and form a continuous trajectory, the iDA embedding may find natural or unnatural breaks in trajectories and the resulting embedding will produce a more discrete structure than actually exists. Because of this, the use of iDA is most appropriate in datasets where the latent structure is thought to be of relatively discrete clusters.

The "pseudo cell" mixture datasets from the CellBench experiments, which includes 9 cell combinations of the same 3 cell lines yielding 34 unique groups with a continuous structure, show the case where cell lines are not discretely separated from each other (Figure 2.9C). The cumulative F-statistic shows that even in the case of continuous, trajectory-like structure between mixes of different cell type compositions, the iDA unsupervised (no number of clusters given as input) and iDA supervised (the estimated number of latent clusters is given as input) either outperforms or matches the performance that a PCA embedding finds with respect to attributing within to between cluster separation between cell type groupings.

In standard pre-processing using PCA, classes which are distinctly defined but have low variation will not be separated on the axes of maximum variation. We have shown that iDA is sensitive to classes with unequal within-class variation and the iDA embedding better separates clusters and identifies the class assignments which

determine a more optimal clustering in the original data space, as evidenced by the Dunn Index. The iDA embedding can be evaluated to determine which features are contributing the most to the class separation and can be used as covariates in downstream analysis to correct for this latent clustering.

2.6 Software Availability

R (http://www.r-project.org) code for the iDA package is available at https://github.com/reesyxan/iDA.

Chapter 3: Introduction to investigating -omics data types in neurons

3.1 Central dogma of biology within neurons

Neurons are a unique classification of cell types in many ways but one of the most notable differences between themselves and homogeneous, symmetric cells is their unique structure. Like other cell types, neurons have a centralized cell body where the nucleus is housed containing each cell's DNA. In addition, neurons possess long projections, dendrites and axons, that enable them to connect to and communicate with neighboring cells, as well as cells that are extremely far away.

With this unique structure comes unique regulatory mechanisms to molecularly support the growth, development, and homeostatic maturity of neurons and their sub-compartments (Fernandopulle et al., 2021; Holt & Schuman, 2013). Neurons transport molecular packages and organelles bidirectionally both from nucleus to axon (anterograde) and axon to nucleus (retrograde) using a series of polarized microtubule tracks with mechanisms known as fast axonal transport and slow axonal transport (Brady et al., 2014; Dalla Costa et al., 2021; Ganguly et al., 2021). The motor proteins kinesin and dynein work together to transport materials in the plus-end-direction and minus-end-direction, respectively, and offer an energetically efficient means for axonal transport and presynaptic replenishment (Guedes-Dias & Holzbaur, 2019).

Fast axonal transport moves packages at a rate of up to 2 micrometers per second. Considering axons can be millimeters to meters long, it can take hours up to days to transport vital proteins, organelles, and other molecules needed to support neuronal functions (Maday et al., 2014). Recent studies have shown that in addition to organelle and protein axonal localization, some neurons localize translationally silenced mRNAs to axonal subcompartments where they can then be translated locally in order to meet the axonal local proteomic requirements during development and maturity (Dougherty, 2017; Glock et al., 2021; Heiman et al., 2008; Ostroff et al., 2019; Sanz et al., 2009; Shigeoka et al., 2016; Tushev et al., 2018). This process of local protein translation greatly reduces the time-sensitive anterograde burden and enables neurons to quickly translate the needed proteins locally to where they will be functionally incorporated.

3.2 Regulation of local protein translation

Axonogenesis and synaptogenesis have been shown to be regulated by differential mRNA shuttling to axonal subcompartments where local protein translation takes place (Preußner & Heyd, 2016; Shigeoka et al., 2016). The local protein translation process can be modulated by extrinsic and intrinsic cues (Jung et al., 2012). An example of the coupling of external growth cues with a shift in local protein translation is beta-actin translation in response to growth signaling cue Netrin-1 where the attractive Netrin-1 causes 3’ untranslated region (3’ UTR)- mediated selective localization of beta-actin within axonal growth cones (Leung et al., 2006; Welshhans & Bassell, 2011; Yao et al., 2006). In parallel, past research has shown

that local protein translation is an activity-dependent mechanism for rapid protein production to direct axon growth during development (Choe & Cho, 2020). In particular, increased neuronal activity leads to an increase in local translation mediated by microRNAs (miRNAs) and RNA binding proteins (RBPs) in the neuropil (Jung et al., 2014; Narayanan et al., 2007; Sutton et al., 2004). Cis-acting miRNAs and RBPs translationally silence localized mRNAs by binding to UTR regions, creating a physical block for translational machinery binding. In the case of beta-actin, past studies have shown that neuronal activity "unmasks" beta-actin mRNA molecules by causing cis-acting UTR-bound elements to disassociate from the mRNA, therefore making them available to ribosomal machinery (Buxbaum et al., 2014). This UTR unmasking was positively correlated with the rate of beta-actin translation. Together, this suggests that shifts in local protein translation caused by a disruption of extrinsic or intrinsic could lead to developmental defects.

3.3 ipRGCs to study activity-dependent local protein translation

Researchers have turned to developing new experimental methods to study the dynamic properties of local protein translation in both developing and mature neural circuits. To study the phenomenon of the dynamic shifts in the translatome populations across development as well as in perturbed activity states, we leveraged the unique properties of intrinsically photosensitive retinal ganglion cells (ipRGCs). ipRGCs are a class of retinal ganglion cells (RGCs), making up just 1%-3% of the total RGC population (Do & Yau, 2010), that express the photopigment melanopsin (Opn4). Unlike conventional RGCs, Opn4 expression in ipRGCs enables them to

respond to light cues independent of rods and cones and control functions such as circadian photoentrainment, pupillary light reflex, and metabolism (Beier et al., 2020; D. Berson, 2003; Hattar, 2002; Preußner & Heyd, 2016; Renna et al., 2011; Berry et al., 2023). ipRGCs are responsible for relaying environmental photic information from the retina to several brain targets and they possess three main properties that provide a model system for studying activity-dependent developmental local translation: 1. ipRGCs have physical separation between cell subcompartments with cell bodies in the retina and long range axonal projections in the brain, 2. ipRGCs have cell type-specific functions and innervation patterns allowing the evaluation of cell-type specific effects, and 3. ipRGCs are genetically tractable and the intrinsic photosensitivity can be exploited to perturb neuronal activity levels. The aggregation of these properties provides an ideal system to study dynamic, activity-dependent translatomic shifts in neuronal cell types.

3.3.1 ipRGCs subcompartments are physically separated

Studying local subcompartment translational regulation is essential for understanding subcompartment-specific changes. In order to study the subcompartment-specific local translation, cell subcompartments (somata versus axons) must be able to be physically separated from one another. ipRGC somata are born in the ganglion cell layer during retina cell type neurogenesis at embryonic days 10-12 (Lucas & Schmidt, 2019; McNeill et al., 2011). They then begin to extend projections during neurite outgrowth to several brain targets including the suprachiasmatic nucleus (SCN) and the dorsal lateral geniculate nucleus (dLGN) where they form long-range

connections from eye-to-brain (Do, 2019). At maturity, these axonal tracts run ~10 mm axon from retina to the SCN and ~15 mm axon from retina to the dLGN (K.-Y. Kim et al., 2019). Because these neurons innervate several brain regions that can be microdissected separately, the translatomes in the somata within the retinas can be harvested separately from axonal compartments in the SCN and dLGN for translatomic comparison.

3.3.2 ipRGCs are defined by functional, developmental, and regional types

ipRGC types have been characterized as modulating a wide array of biological processes, such as regulation of sleep/awake cycles, metabolism, hunger patterns and hormone production (Lazzerini Ospri et al., 2017). ipRGCs form parallel pathways to several brain targets to photoentrain these functional roles. ipRGCs projecting to the brain's master circadian pacemaker, the suprachiasmatic nucleus (SCN) play a pivotal role in coordinating the circadian rhythm to photic input, while ipRGCs projecting to the dorsal lateral geniculate nucleus (dLGN) play a role in in visual perception.

By postnatal day 0 (P0), ipRGC axons have significantly innervated multiple visual areas of the brain, including the dLGN (McNeill et al., 2011). However, this is not the case for the SCN. Although ipRGC axons pass by the SCN, they do not begin innervating the structure for several days (and do not finish doing so for several days more). The ipRGCs that innervate the SCN are known to be M1 ipRGCs, and even more specifically, are a molecularly defined subset of M1 ipRGCs because they lack the expression of Brn3b (Chen et al., 2011; Do, 2019). M1 ipRGCs innervate the

SCN are developmentally delayed when compared to non-M1 ipRGCs innervating other areas of the visual brain including the dLGN. Based on these findings, ipRGC types are developmentally and regionally unique; SCN-projecting M1 ipRGCs have a different developmental timeline than dLGN-projecting non-M1 ipRGCs.

3.3.3 ipRGCs are genetically tractable

ipRGCs express the photosensitive protein melanopsin (Opn4). ipRGC axon development and early synaptogenesis in central targets occurs during a period where the photosensitivity of the retina is determined solely by the presence of the photosensitive protein melanopsin (D. M. Berson, 2002; Fu et al., 2005; Hattar, 2002; Provencio et al., 1998; Renna et al., 2011). Disruption of melanopsin-driven phototransduction has been shown to be critical in maintaining developmental features including retinal wave properties, retinal vasculature development, and retinofugal projection development (discussed further in chapter 5) (Renna et al., 2011; Tiriac et al., 2018). In addition, the melanopsin phototransduction pathway also drives behavioral and developmental defects such as neonatal photoaversion and synaptogenesis (Caval-Holme et al., 2022; Hu et al., 2022). Particularly, melanopsin-driven phototransduction in ipRGCs directly affects synaptogenesis in various brain targets by way of oxytocin release in the supraoptic nucleus and the paraventricular nucleus (PVN) (Hu et al., 2022). The direct link between melanopsin-driven activity in ipRGCs during development and intracellular translational regulation driving such developmental features has yet to be explored. Because of this evidence of the

regulation of local translation by neuronal activity, melanopsin expression dictating action potentials evoked by photic stimuli may play an important role in regulating activity-dependent plasticity of local protein production in the development of ipRGC circuits. Since Opn4 is both a unique cell type marker and controls a whole host of phototransduction-dependent developmental features, it is an ideal candidate to study ipRGC-specific transcriptomics and translatomics.

3.4 Neuronal Translatomes: Axon-TRAP-RiboTag

Evaluating different regulatory processes throughout the central dogma pipeline from DNA to protein is critical in understanding developmental and homeostatic neuronal processes. Tight regulation of transcription and translation allows for not only cell type functional specificity, but also cell compartment functional specificity, especially in systems where transcription and translation occur in distant compartments of the cell, such as neural tissue. Specifically, the translational regulation within cell compartments allows for complexity and precision of individual cell compartments. Studying the neuronal compartment translatomes has become vital to investigate axonal regulatory properties by quantifying the local protein translation in axons that enables rapid responses to cell-cell signaling cues during development as discussed in previous sections. Axon-TRAP-RiboTag, is a new experimental method developed by the Holt Lab to quantify the translatome in axons (Shigeoka et al., 2018). RiboTag is a genetic approach which allows for immunoprecipitation of ribosome bound mRNA in the presence of knocked-in Cre driver by labeling ribosomes with an HA epitope. RiboTag consists of a knocked in floxed wildtype

Rpl22 exon4 followed by an Rpl22 exon 4 which has an inserted HA epitope. When crossed with a mouse line containing a Cre recombinase driver, cells which express the Cre recombinase will also express the Rpl22^{HA} mutant (Figure 3.1). The HA positive ribosomes can then be isolated by way of immunoprecipitation using an antibody for HA.

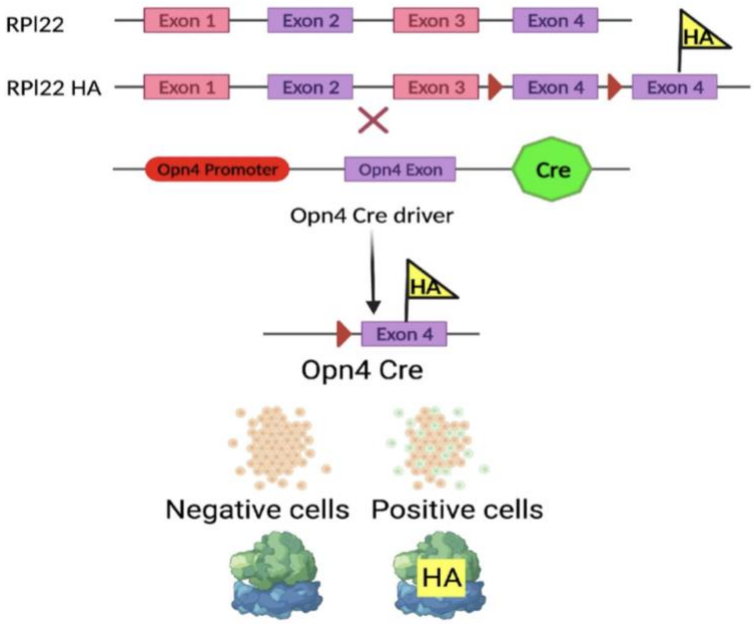


Figure 3.1. RiboTag genetic approach with Opn4 Cre driver.
RiboTag consists of a knocked-in floxed wildtype Rpl22 exon4 followed by an Rpl22 exon 4 which has an inserted HA epitope. When crossed with a mouse line containing a Cre recombinase driver (such as Opn4 shown in this figure), cells which express the Cre recombinase will also express the Rpl22^{HA} mutant (green cells in bottom row). Cre negative cells will not have HA labeled ribosomes (orange cells in bottom row).

In the case of ipRGCs, melanopsin mutant Opn4^{CRE} transgenic mice (Ecker et al., 2010) were crossed with the transgenic mouse line Rpl22^{HA} (Sanz et al., 2019) which has an HA epitope tag embedded in the coding region of the large ribosomal subunit

protein (illustrated in Figure 3.1). This cross produced a colony of mice that generates HA epitope-tagged ribosomes only in the presence of Cre recombinase. In the case of $Opn4^{CRE}$ x $Rpl22^{HA}$, the Cre recombinase is expressed in cells that express Opn4, labeling ribosomes in ipRGCs. Ribosome-bound mRNAs in tissue collected from these mice can then be isolated using an anti-HA antibody, effectively isolating the ipRGC translatome, and mRNAs then sequenced using low-input RNA-seq protocols.

3.5 Contributions

New experimental protocols are pushing the bounds of attainable information explaining how cells develop, function, and maintain homeostasis. Experiments using complex, cutting-edge sample collection and isolation protocols like axon-TRAP-Ribotag require careful computational exploration to validate the data. Particularly when experiments require several experimental batches, understanding sources of variation (unwanted and wanted biological and technical contributions) is extremely important to ensure validity of downstream results. A challenge in axon-TRAP-Ribotag experiments is consistency of the isolation of the translating mRNA population from the bulk mRNA population. There is a need for metrics to evaluate the efficiency and consistency of the protocol across experimental and technical variables. Only after potential batch effect contributors are identified and eradicated by sample filtering or batch correction methods can downstream analysis like differential expression analysis or gene set enrichment analysis be done.

In chapter 4, we provide a case study of special considerations for a complex, multi-batch high throughput experiment. We investigate the multi-faceted heterogenic contributions of a study using the axon-TRAP-RiboTag translatomic isolation protocol in a neuronal cell type. Our results highlight that batch-correction methods may reduce signal due to true biological heterogeneity in addition to technical noise. Further, we show axon-TRAP-Ribotag experiments are subject to varying degrees of isolation efficiency with less-efficient pulldowns having a closer resemblance to bulk RNA-seq. We offer metrics to help identify biologically signal-driven batch-differences specific to translatomic experiments.

In chapter 5, we employ our understanding of variational contributions from chapter 4 in the intrinsically photosensitive retinal ganglion cell (ipRGC) -omics case study to explore the biological transcriptomic and translatomic coordination. Our analysis revealed ipRGCs participate in subcompartment-specific local protein translation. Genetic perturbations of the Opn4 photopigment-driven neuronal activity led to global tissue transcriptomic shifts in both the retina and brain targets, but the ipRGC axonal-specific translatome was unaltered.

Chapter 4: Validating the efficacy of Axon-TRAP RNA-Sequencing Results in Multi-Batch Studies

4.1 Background

4.1.1 Multi-batch processing in high-throughput studies

Since their introduction, bulk RNA-sequencing studies have helped experimentalists better understand steady-state RNA levels and have become increasingly more accessible. In order to expand the range of biological information that can be obtained beyond RNA steady-state levels, experimental protocols have grown more complex enabling quantification of spatiotemporally precise expression regulation. In cases where spatial regulation can be studied in cell compartments, such as in long-range neuronal connections, to relate cell type soma and axonal compartment's regulation, innovative experimental methods are being developed. Scientists in the field of neurobiology have turned to methods such as translating ribosome affinity purification (TRAP) and axon-TRAP-RiboTag to quantify the translatome, a post-transcriptional state, to uncover complex biological phenomena (Reynoso et al., 2015; Shigeoka et al., 2018). These studies frequently involve a variety of experimental conditions, including various tissue locations, genotypic profiles, and developmental ages, requiring multiple biological replicates to be collected. Careful consideration of the experimental design needs to be taken since a large number of samples are needed to guarantee sufficient statistical power for subsequent analyses (Auer & Doerge, 2010). Also, tissue collection techniques frequently call for pooling of tissue samples due to the limited amount of starting tissue available, particularly in the axonal

compartments of the cells, further increasing the number of samples needed for a study.

These studies are frequently processed in multiple batches due to limitations on the number of samples that can be collected, processed, and sequenced with staffing and time limitations. To eliminate erroneous results in subsequent analyses, it is necessary to deconvolve potential batch effect (variation due to non-biological factors in an experiment) driving variational contributions in complex experiments with numerous experimental conditions as well as when processing or experimental batches (Sheerin et al., 2019). Particularly in experiments using novel, state-of-the-art sample collection and isolation protocols, it is of the utmost importance for investigators to do the due diligence to investigate all intended or unintended variables which have the potential to confound downstream results.

4.1.2 Assumptions made by batch-correction methods may not often be met

Deciphering the true biological signal from technical noise in these data is necessary but difficult. Many batch correction methods have been developed to adjust count data so that variational contributions due to technical noise are minimized, such as surrogate variable analysis (SVA), COMBAT-seq, limma, and RUV (Leek & Storey, 2007; Risso et al., 2014; Ritchie et al., 2015; Y. Zhang et al., 2020). SVA estimates latent surrogate variables (SVs) to model variation not attributed to condition of interest (i.e. unknown/unmodeled sources of unwanted variation) such that the SVs can be used as adjustment variables in downstream models for differential expression (DE) analysis, for example. COMBAT-seq models variation attributed to a measured

batch variable explicitly (i.e. known sources of unwanted variation). Each of these methods is extremely powerful but make some assumptions which may not be true without further investigation. If these assumptions are not valid, downstream results will not be statistically valid. For example, with the SVA method the assumption is that all unknown/unmodeled sources of variation are unwanted, while the COMBAT-seq method assumes that the modeled batch variable accounts for all unwanted variation. If untrue, the results from using these methods to adjust for noise in RNA-sequencing data will cause either an increase in type 2 errors (false-negative results), if adjustment variables adjust for variation correlated with true biological variation, or an increase in type 1 errors (false-positive results) if adjustment variables do not adequately capture technical noise. Consequently, it is essential to assess model assumptions before using batch correction methods to ensure the accuracy of the downstream results.

4.1.3 Case Study of axon-TRAP RNA-seq data from long range neuronal connections from eye-to-brain

I aimed to study the translational changes in a specific retinal ganglion cell, the intrinsically photosensitive retinal ganglion cells (ipRGCs) across developmental age, different mouse genotypes, and in two cell compartments (soma versus axonal compartments). This study was aimed at testing the hypothesis that signaling driven by a particular gene's expression (melanopsin) regulates local protein translation in developing ipRGCs. Using a knock-in mouse that expresses Cre-recombinase in place of endogenous melanopsin ($Opn4^{Cre}$), we labeled ribosomal proteins with a floxed

$Rpl22^{HA}$ line ($Rpl22^{HA}$ flanked by loxP sites) and developed an ipRGC-TRAP assay (as described in Section 5.6.1). This assay allowed us to isolate ribosomes and sequence the associated translating mRNAs from ipRGC cell bodies in the retina and axons in the dorsal lateral geniculate nucleus (dLGN) and suprachiasmatic nucleus (SCN). We performed these experiments before (P8) and after (P15) eye-opening for comparison with adult (P60) translatomes in both control and melanopsin knockout mice ($Opn4^{Cre/Cre}$::$Rpl22^{HA/+}$).

This study involved three experimental variables for which I wanted to assess translatomic changes in: developmental age, Opn4 genotype, and sample collection location. Due to the complex nature of this experimental setup, several sources of potential unwanted technical variation were practically unavoidable, including two experimentalists collecting and preparing samples (referred to as Experimentalist A collecting samples in batch 1 and Experimentalist B in batch 2) and two library preparation and sequencing sources (experimental batch 1 performed commercially, experimental batch 2 performed at the University of Maryland BBI-AGTC). A total of 81 samples were collected and sequenced for this study, 61 samples in experimental batch 1 and 23 samples in experimental batch 2.

Principal component analysis (PCA) computed for all samples surviving sample filtering (n = 78) revealed discrete clustering in both the PC1 (36.09% variance) and the PC2 (21.74% variance) projections (Figure 4.1). When visualized with the samples colored and shaped according to the experimental variables of interest

(sample location = color hue, genotype = within-color shade, developmental age = shape) (Figure 4.1, A), it is apparent that PC1 variance is attributed to separation of retina samples (pinks) versus brain region samples (blues and greens), but there is no known biological variable attributed to the clustering in the PC2 projection). Clustering in the PC2 dimension is clearly attributed to the experimental batch (presumed unwanted variance) after visualizing the top two PCs according to the experimental batch (Figure 4.1, B). This discreteness caused by technical variables (experimentalist changes as well as library preparation and sequencing changes) would lead an investigator to believe that there is a need for a batch correction before conducting any downstream analysis like DE or GSEA to ensure the validity of the results.

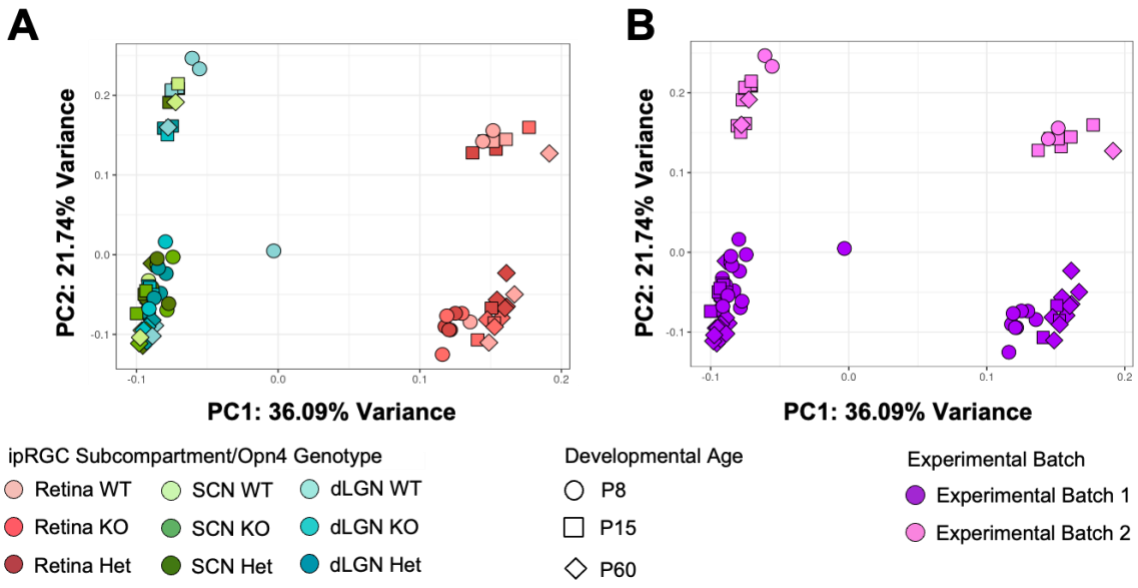


Figure 4.1. Batch effect in ipRGC case study.
Principal component analysis (PCA) of the log2 CPM transformed counts normalized with quantile normalization computed for all samples (n = 78). PC1 accounts for 36.09% of the total variance and PC2 accounts for 21.74% variance of the total variance. The PCA plot is labeled based on different meta factors: A) Samples labeled by ipRGC subcompartment location/genotype/developmental age. B) Samples labeled by experimental batch.

With an apparent need for batch correction, researchers are left choosing between batch correction methods such as SVA or COMBAT-seq, but discrepancies in resulting "batch corrected" count data from different methods beg the question of which method is appropriate for the data at hand. For example, in the axon-TRAP ipRGC dataset, the SVA method resolves location-specific variability in the PCA projections from the SV corrected dataset, but there does not appear to be large variational differences due to genotype or developmental age as expected (Figure 4.2). In contrast, the COMBAT-seq batch-corrected counts resolve location and developmental age specificity, but not genotype. The discrepancies between the two

batch-correction methods suggest that there is a large contribution of variability attributed to sources that are not the measured biological variables of interest (based on the SVA correction), but this portion of variance is not wholly modeled by the experimental batch variable (based on the COMBAT-seq correction). This discrepancy raises the need for an evaluation of biological and technical variability to ensure that there are no unmeasured/unmodeled biological or technical sources of variability playing a role in batch-driven sample differences. To address this, I propose an exploratory pipeline including several diagnostic metrics to parse sources of variability within a high-throughput sequencing dataset. I find that variational contributions which are often assigned as "unwanted noise" are actually from true biological signal that, if corrected for using batch correction methods, can lead to spurious results.

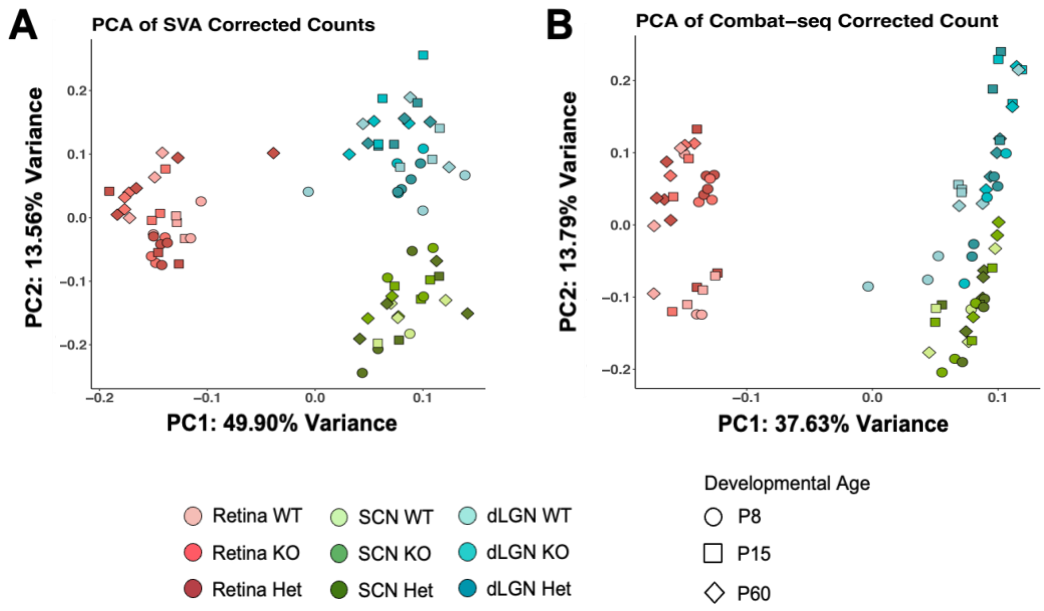


Figure 4.2. SVA vs COMBAT batch-correction.
PCA of the batch-corrected counts by two popular methods. A) Surrogate Variable Analysis (SVA) was conducted using location/genotype/developmental time as the input variable with respect to which the variation is preserved. B) COMBAT-seq was conducted using the experimental batch variable as input. The plot is labeled by ipRGC subcompartment location/Opn4 genotype/developmental age.

4.2 Results

4.2.1 Experimental batch contributes to unwanted batch effect

I aimed to distinguish if the measured experimental batch variable contributed to any unwanted variability between biological sample groups of interest. I employed surrogate variable analysis (SVA) to estimate the contribution of sources of variation not attributed to the known variables of interest (location, genotype, and time of development). In using these biological variables of interest, I defined the full model matrix as:

$$\underbrace{expression_i}_{\text{expression of gene i}} = \underbrace{B_0 + B_1 location : genotype : time}_{\text{variables of interest}}$$

and the null model including only an intercept equal to 1. SVA estimates 9 SVs given this full model. Each SV was then modeled as a function of each known source of variation (experimental batch, location, age, and genotype). The F-statistic for each model were then calculated. In SV dimensions 1 and 4 the defining variable contributing to the batch correction estimate is experimental batch (Table 4.1).

Table 4.1. F-statistic for each SV dimension resulting from SVA.

	Batch	Subcompartment Location	Opn4 Genotype	Developmental Age
SV1	6.16	0.07	0.22	0.25
SV2	0.37	0.47	0.10	0.06
SV3	0.13	0.14	0.04	0.04
SV4	1.01	0.33	0.02	0.19
SV5	0.02	0.02	0.00	0.01
SV6	0.03	0.06	0.02	0.06
SV7	0.01	0.00	0.05	0.05
SV8	0.00	0.01	0.01	0.09
SV9	0.01	0.02	0.03	0.05

The SV contributing the majority of the batch correction attributed to the experimental batch is SV1, but there is a small amount of variation attributed to

biological variables of interest (location, genotype, and developmental time) (Figure 4.3). These small contributions attributed to biological variables of interest were likely due to confounding between the experimental batch and genotype (Cramer's V test = 0.4817, *P* value = 0.0001) and developmental age (Cramer's V test = 0.4972, *P* value = 0.00006).

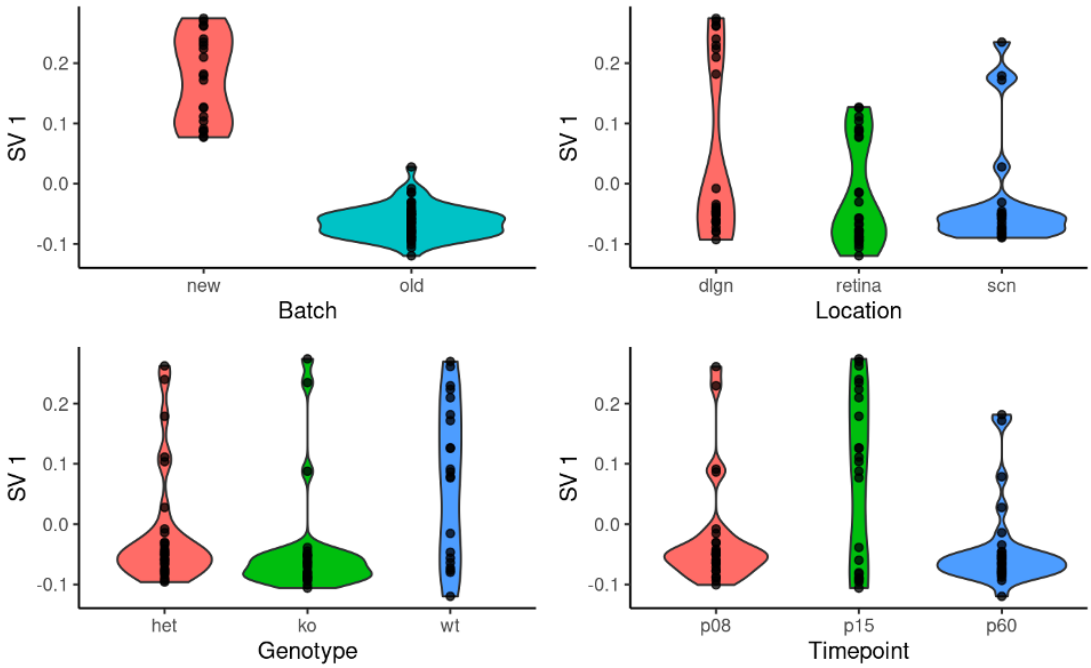


Figure 4.3. SVA versus known experimental factors.
Distribution of Surrogate Variable 1 (SV1) values across samples versus each variable known to contribute variability to the translatomic data. The experimental batch (top left panel) dominates the batch correction for SV1, while the variables of interest (location, genotype, and developmental time point) have variability preserved.

4.2.2 Experimental batch sequencing depth does not correlate with unique genes observed

Although the total sequenced read counts were consistent across all samples in both experimental batches, the experimental batches in the axon-TRAP ipRGC dataset display an imbalance in the total *mapped* read counts due to differences in read count mapping rates (the rate at which the sequenced reads successfully map to the reference genome). Experimental batch 2 samples resulted in a much lower mapping rate for sequenced reads mapping to coding sequence regions (and had higher mapping rates for reads mapping to rRNA regions) when compared to experimental batch 1, meaning the ribo-depletion in experimental batch 2 was not as efficient as experimental batch 1. Experimental batch 1 (n = 57) has an average of 18,239,115 coding sequence-mapped read counts per sample and Experimental batch 2 (n = 21) has an average of 9,835,858 coding sequence-mapped read counts) (Figure 4.4).

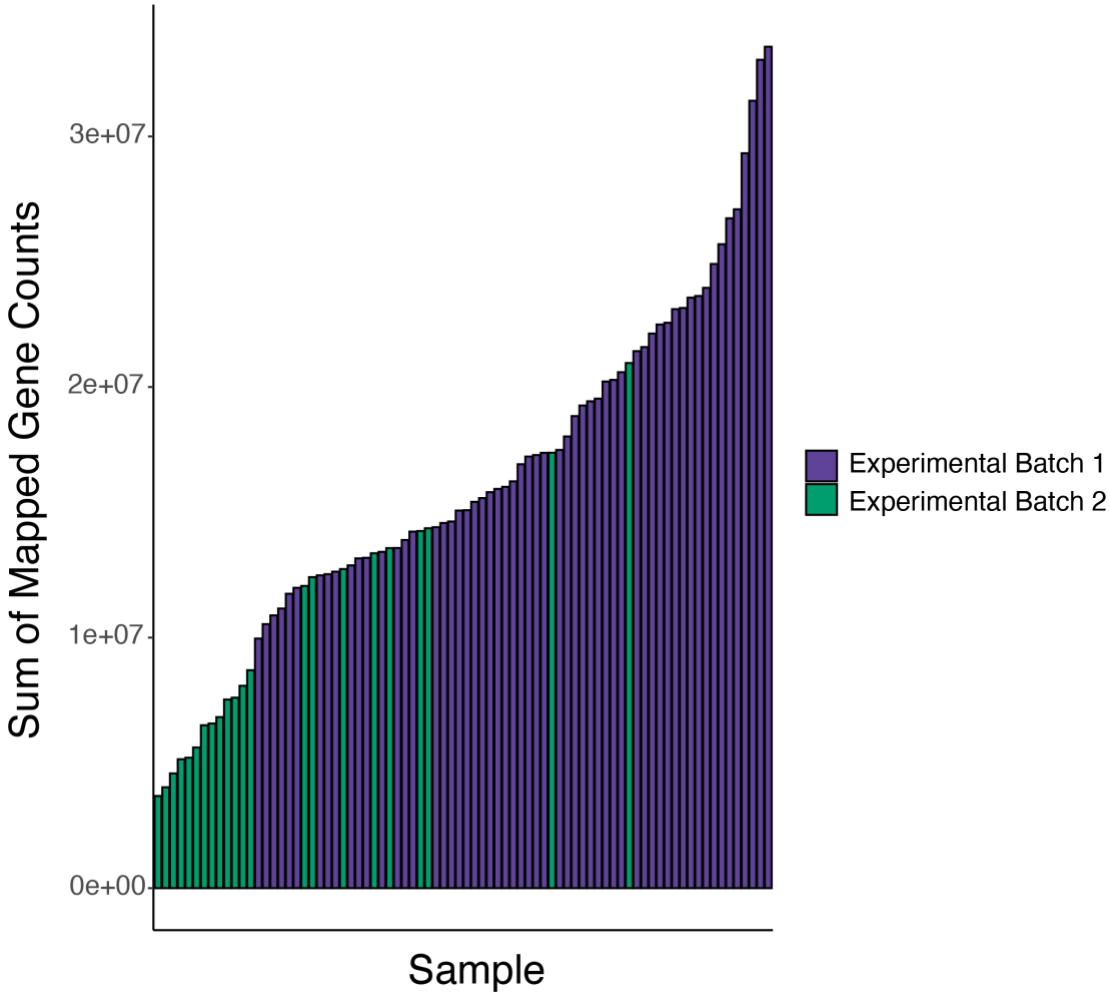


Figure 4.4. Mapped read counts per sample.
Sum total number of mapped read counts to the mouse genome (GRCm38) per sample. Samples collected in Experimental Batch 1 are colored purple, samples collected in Experimental Batch 2 are colored green.

Evaluating the number of genes with detectable expression levels, referred to as the number of non-zero genes observed, per sample versus the observed normalized number of mapped reads showed an apparent relationship between these two variables (Figure 4.5). The subset of samples which have lower non-zero genes which are predominantly experimental batch 2 samples also have lower sequencing depth

(observed as mapped reads). This led us to address the question: are the lower numbers of genes observed in experimental batch 2 only attributable to lower depth of sequencing (total number of reads which were successfully mapped to the reference genome)?

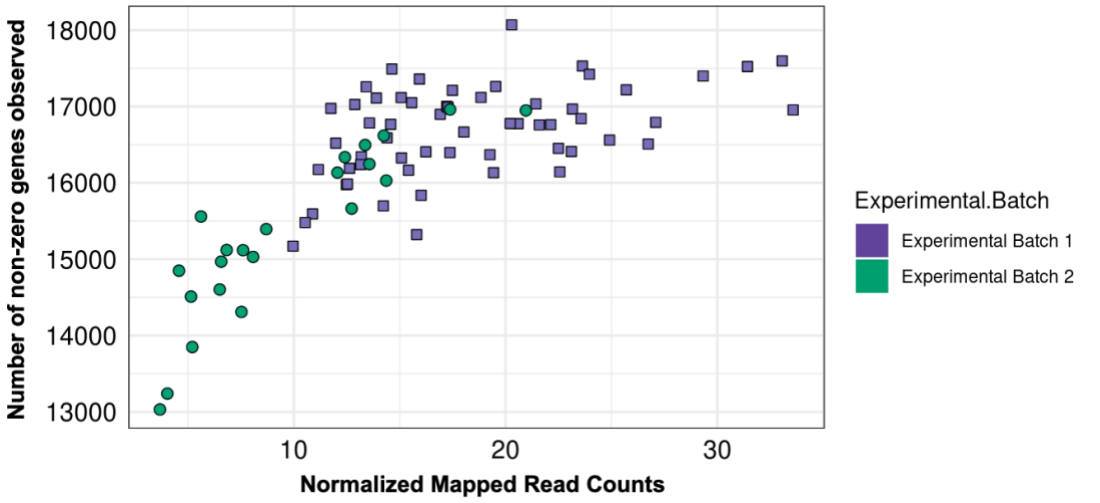


Figure 4.5. Relationship of mapped reads versus genes observed.
The number of normalized mapped reads versus the number of non-zero genes observed (genes with counts over 1 were counted as being observed) per sample. The relationship between the two reaches an asymptotic limit, with the maximum number of genes observed in an individual sample being just over 18,000 genes.

To test if the differences in experimental batches are solely due to transcriptome sequencing depth, I downsampled the libraries for samples in experimental batch 1 to match the library sizes of those samples in experimental batch 2 as described in Section 4.5.5: Sequencing Depth Contributional Assessment. Following the bootstrapped downsampling for each experimental batch 1 sample (n = 57, purple), the number of non-zero genes observed did decrease to a dynamic range of ~ 14,900 - ~ 17,900 compared to ~ 15,100 - ~ 18,050 before downsampling (Figure 4.6).

Though there was a reduction, the number of non-zero genes observed per sample still did not resemble that of the experimental batch 2 biological samples based on visual inspection (n = 22, green) (Figure 4.6).

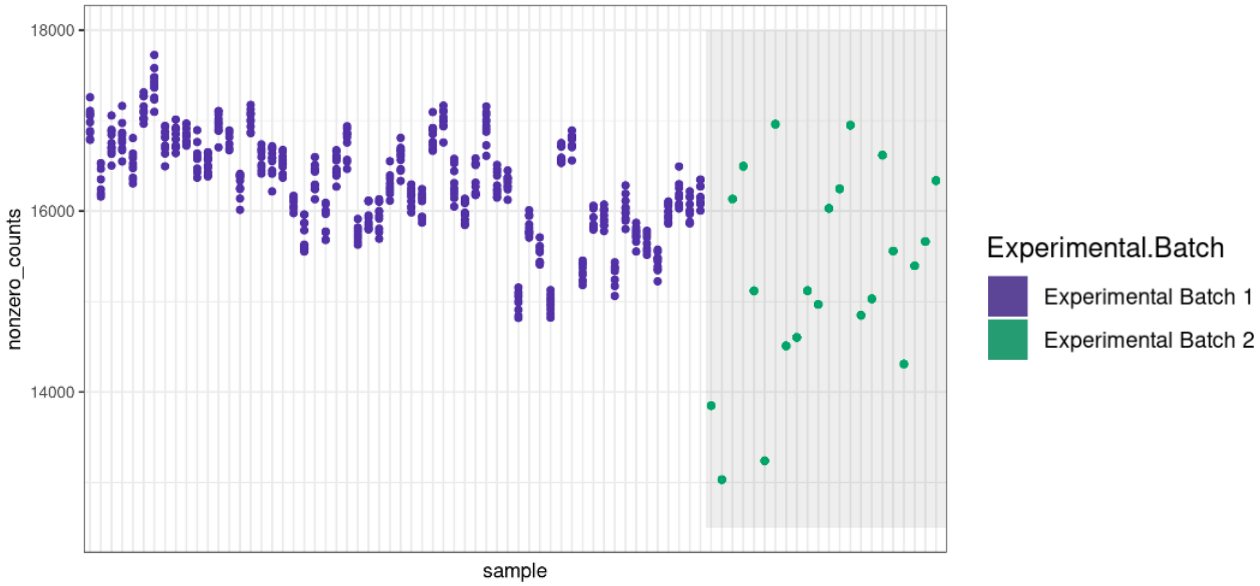


Figure 4.6. Bootstrapped downsampling of sequencing counts within each sample.

Number of genes observed in each sample after downsampling (y-axis) per sample (x-axis). Bootstrapped downsampling for each sample's gene counts was conducted. For each sample, the gene counts were sampled with the probability of sampling a gene count being proportional to the gene's observed expression. Downsampling was repeated 10 times per sample in experimental batch 1 (purple).

4.2.3 Batch effect can be driven by true biological signal

To evaluate if the difference in the number of non-zero genes observed was due to an unintended source, I visualized the number of non-zero genes observed per sample by indicating two potentially confounding biological variables: genotype and sample location (Figure 4.7). The color and shape of each point indicates two measured

biological variables; each point is colored by location and the shape indicates the genotype. The samples in Experimental batch 2 (shaded with gray background on the right-hand side) do not show differences in the number of non-zero genes observed with respect to genotype (shapes), but there are distinct differences between the number of non-zero genes due to location (shaded background). In experimental batch 2, samples isolated from the retina, i.e. samples of neuronal cell bodies, (pink) have a larger number of unique mRNA species observed than their axonal sample counterparts (SCN - green, dLGN - blue). The experimental batch 1 samples (unshaded background) show no clustering based on either genotype or location, particularly in samples collected earlier in the study.

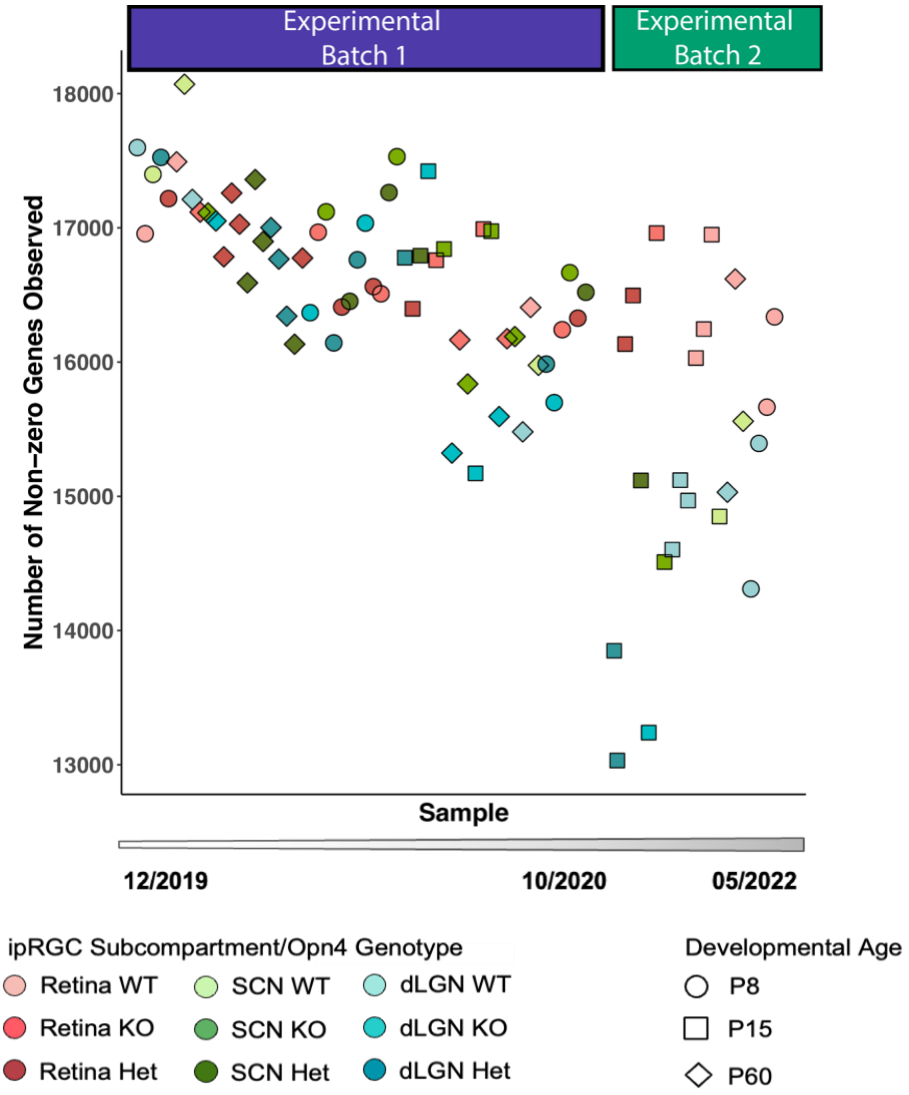


Figure 4.7. Non-zero genes quantified per sample by subcompartment location/genotype.

Number of non-zero genes observed in experimental batch 1 (purple region, samples collected by experimentalist A from 12/2019 to 10/2020) and experimental batch 2 (green region) (collected by experimentalist B and processed as one batch in 05/2022). The shape of each point indicates the sample genotype and the color indicates the cellular region of each sample.

To understand the efficiency of the axon-TRAP-RiboTag isolation of ribosome-bound mRNAs from the population of all mRNAs within the collected tissue, we performed bulk RNA-sequencing of each tissue location (retina, SCN, and dLGN) for

comparison. The number of non-zero genes per sample processed by bulk RNA-sequencing were then computed (Figure 4.8). The 15 bulk RNA-sequenced samples collected from the 3 locations revealed a similar number of non-zero genes observed to samples collected in experimental batch 1 (Figure 4.8, purple region = experimental batch 1 samples, black shaded region = bulk RNA-seq samples). The bulk RNA-seq samples showed an average of ~17,500 genes observed among all samples (black region), regardless of tissue location in contrast to the location-specific number of non-zero genes observed in experimental batch 2 (green region).

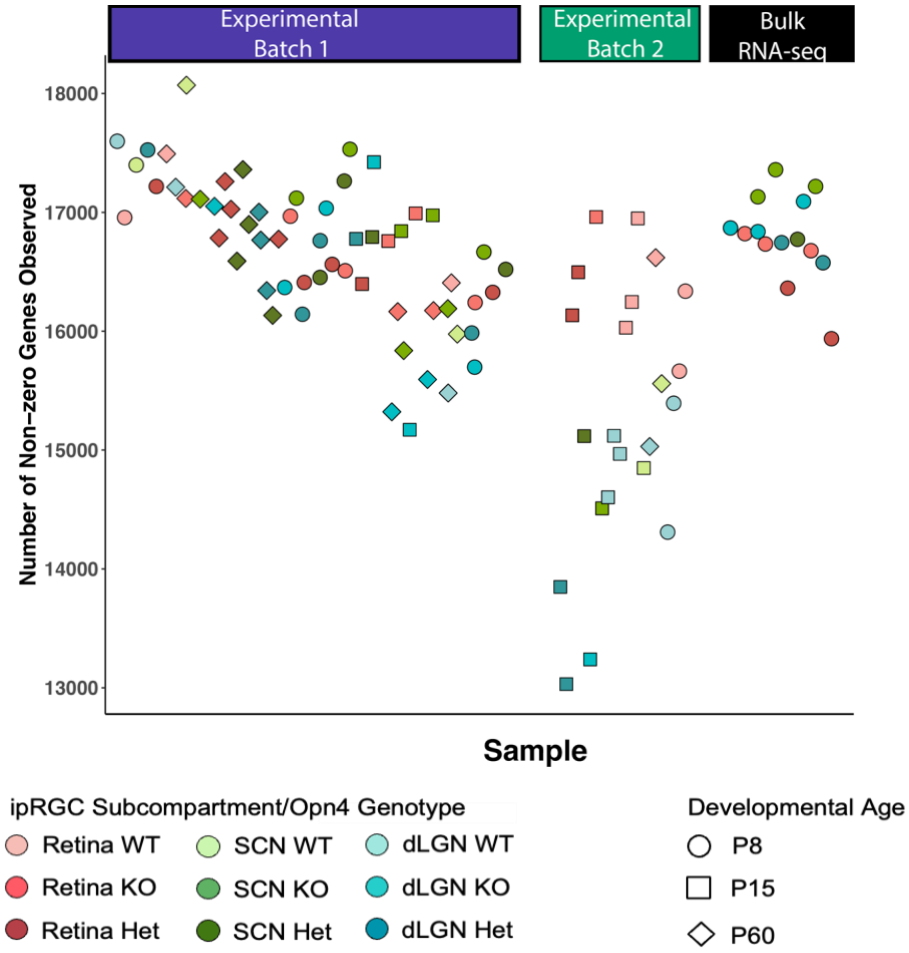


Figure 4.8. Non-zero genes observed between axon-TRAP and bulk RNA-seq.

Comparison of the number of non-zero genes observed between axon-TRAP-RiboTag samples (experimental batch 1—purple region, experimental batch 2—green region) and bulk RNA-sequencing samples (black region).

I followed the same workflow to compare the number of non-zero genes observed in the samples from another axon-TRAP-RiboTag study with a similar experimental design (Shigeoka et al., 2016). Shigeoka et al. collected ribosome bound mRNAs from soma and axonal compartments in retinal ganglion cells (RGCs) projecting from the retina to the superior colliculus (SC) in two genotypes across four developmental ages. Consistent with our findings, no apparent association between differences in the

number of non-zero genes observed due to developmental age (shape of point) or genotype (size of point) were found, but the axonal SC samples do have fewer non-zero counts compared to the cell body retina samples (color) (Figure 4.9).

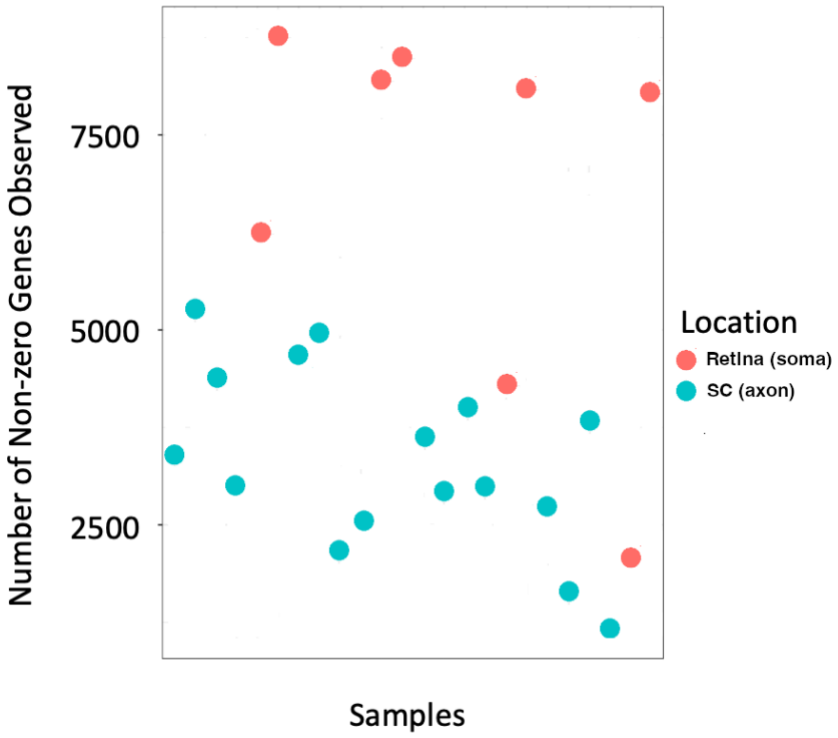


Figure 4.9. Non-zero genes observed in the original axon-TRAP-RiboTag.

Number of non-zero genes (genes with more than 0 counts in an individual sample) observed from samples collected from the retina (neuronal cell bodies) as well as the superior colliculus (SC - axonal compartment) using the axon-TRAP-RiboTag collection protocol from the Holt lab (Shigeoka et al., 2016). Translatomic populations in samples from neuronal cell bodies (pink) have on average higher numbers of non-zero genes represented than axonal samples collected (blue).

4.2.4 Ribosomal protein expression is variable between experimental batches

Variance partition analysis yielded both global views of variance attributed to each experimental variable (both biological and technical) as well as a gene-level view

(Figure 4.10). At the global level across all genes, variance partition analysis revealed that both the biological variable sample location (location, pink) as well as the technical variable experimental batch (green) accounted for a large portion of variability in the ipRGC axon-TRAP data, while the genotype of the samples did not (genotype, blue) (Figure 4.10A). Evaluating the top genes with a majority of the expression heterogeneity attributed to the experimental batch revealed that five out of the top ten genes are ribosomal proteins (Rp) (Figure 4.10B).

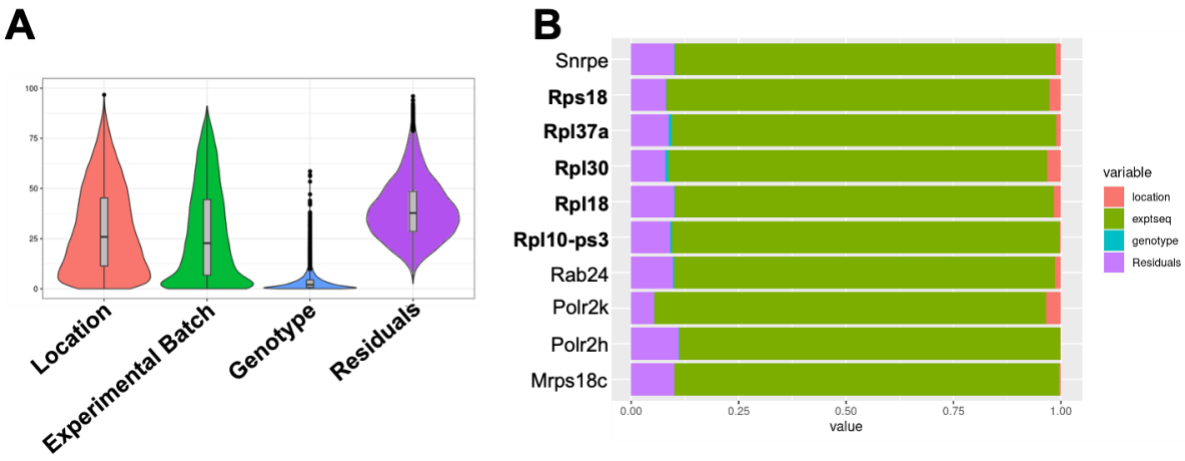


Figure 4.10 Variance partition analysis of axon-TRAP-RiboTag samples.

Variance partition analysis of the axon-TRAP-RiboTag samples collected at one developmental age, P15, to describe the variance explained by both technical and biological variables. A) Biological variable sample location as well as technical variable experimental batch account for a majority of the modeled variability, recapitulating results visualized in the PCA. B) Variance per gene attributed to each input variable (Location (pink), Experimental Batch (green), genotype (blue), and residuals (purple) ranked by variance attributed to experimental batch.

I evaluated the expression of all ribosomal proteins within both experimental batches of samples. The normalized expression values of all ribosomal proteins were hierarchically clustered across all samples. The heatmap of the clustered expression profiles of all ribosomal proteins reveal discrete clustering based on the experimental batch (Figure 4.11).

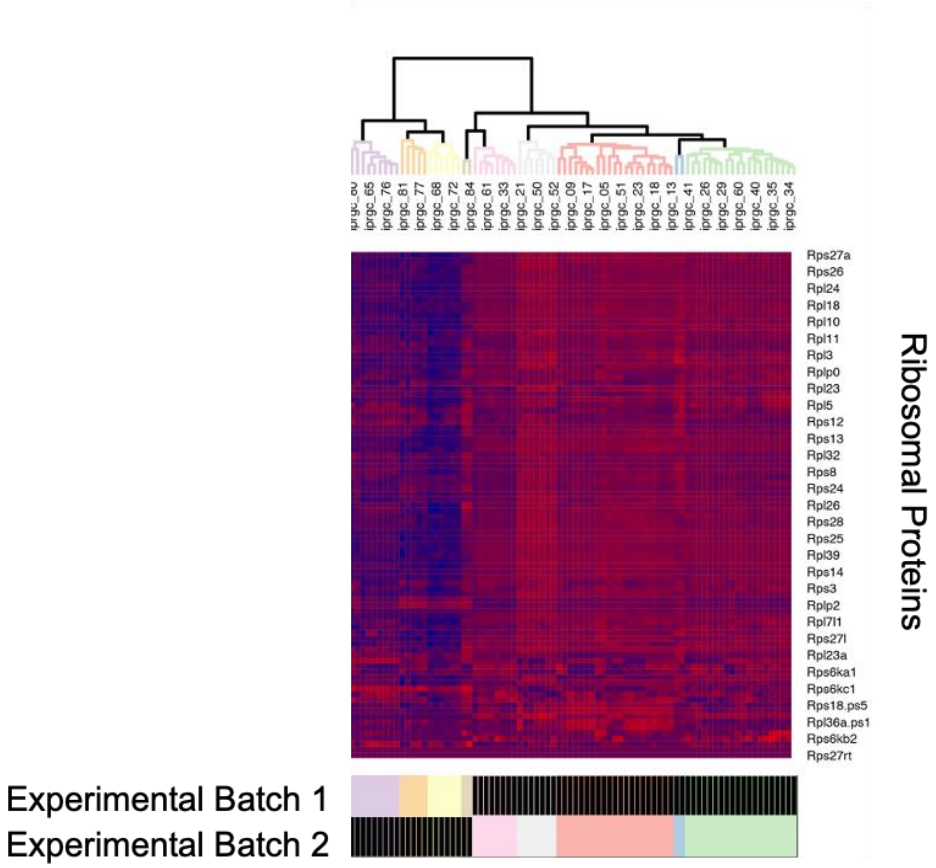


Figure 4.11. Hierarchical clustering of the expression profiles for all ipRGC axon-TRAP samples.
Unsupervised hierarchical clustering of samples based on the normalized expression (log2 CPM transformed counts normalized with quantile normalization) from all ribosomal proteins (Rps, Rpl) in the axon-TRAP dataset.

4.3 Discussion

Multi-batch experiments are often unavoidable due to technical limitations of experimentalists, time constraints, and the volume of samples needed for adequately powered analyses. Samples produced in multiple batches often lead to “batch effect”, i.e. heterogeneity in the data attributed to known or unknown sources of technical, or non-biological, variables (Figure 4.1). It is of principal concern to parse wanted biological variability (variables built into the experimental design), unwanted technical variability (variables often forced to be part of the experimental design), and unwanted/unintended biological variability (biological variables not in the experimental design). This deconvolution requires particular considerations for multi-batch TRAP and axon-TRAP-RiboTag studies, notably ensuring the isolation efficiency is consistent across experimental batches when isolating ribosome-bound mRNAs from free mRNAs - a potential source of unwanted *biological* variation.

If discrete clustering correlated with a measured technical variable such as experimental batch is apparent in dimensionality reduction methods like PCA, investigators often turn directly to batch effect correction methods such as SVA or COMBAT-seq (Figure 4.2). Often, the invocation of these methods is done without exploratory analysis of the source behind batch-driving factors. Particularly in experiments using new experimental sample collection protocols, blindly turning to batch-correction methods in an attempt to preserve heterogeneity due to biological

features of interest is not advisable. If the quelled signal supposedly due to batch effect is in fact heterogeneity due to true underlying biological differences, batch-correction methods will result in underpowered and erroneous downstream results. It is of utmost importance to fully interrogate potential biological and technical differences between batches before proceeding with these methods.

One commonly overlooked but insightful metric in high-throughput studies which can help understand the underlying similarities and differences in multi-batch transcriptomic or translatomic sequencing studies is the number of non-zero genes observed (that is, the number and identity of unique genes detected per sample). The level of transcriptomic/translatomic diversity can reflect both biological variability in gene expression as well as technical effects from sequencing pipelines. If variability exists between the number of non-zero genes observed per samples or per experimental conditions, it is essential to determine if this is a result of biological differences between samples or technical differences in experiments where the sequencing depth is not uniform across samples sequenced (commonly in multi-batch experiments).

An imbalance in sequencing depth (or the number of reads which survive trimming and mapping to the genome for the final counts table) can result in a batch effect, particularly for genes which are not highly expressed and as the sequencing depth increases, the coverage will also increase, with more reads attributed to genes with lower levels of expression. If a biological sample is sequenced such that there is not

adequate coverage, this means there is a lower probability of lowly-expressed genes being sequenced, and therefore they will not be represented in the final count data (Bass et al., 2019). This can lead to misleading apparent differences in gene counts which are purely associated with technical differences and not biological ones.

To test if this imbalance in library size was contributing any technical noise in this axon-TRAP dataset, I downsampled the counts from biological samples with more sequenced reads surviving read trimming and mapping (experimental batch 1) to compare to biological samples with fewer resulting reads (experimental batch 2). Downsampling did not recapitulate the non-zero gene counts found in experimental batch 2, indicating that the discrepancies in the non-zero gene counts observed were not due to technical differences in sequencing depth (Figure 4.6), consistent with previous findings (Bush et al., 2018).

After determining the observed batch effect between experimental batches was not driven by technical differences such as sequencing depth, I turned to evaluating other potential axon-TRAP-RiboTag specific biological confounders. I observed differences in the number of unique mRNA species represented in samples with respect to their location in experimental batch 2, but not in experimental batch 1 (Figure 4.7). The experimental batch 1 samples show no clustering based on either genotype or location, particularly in samples collected earlier in the study, indicating innate, unintended biological differences between experimental batches. Previous studies have found the axonal local translatomes contain fewer mRNA species with

specific axonal functional roles (Glock et al., 2021; Rozenbaum et al., 2018; Shigeoka et al., 2016). In order to decipher the expected distribution of unique ribosome-bound mRNAs in ipRGC cell-compartments from axon-TRAP-RiboTag collection techniques, I compared our results from each experimental batch to those of a comparable axon-TRAP-RiboTag study conducted by Shigeoka et al. (Shigeoka et al., 2016). This study quantified the ribosome-bound mRNA populations in retinal ganglion cells (RGCs) innervating the superior colliculus (SC) in RGC cell bodies (collected from the retina), and RGC axonal compartments (collected from the SC). This study also collected samples at several developmental ages (E17, P1, P7, Adult), and in two genotypes. The discreteness of the quantified number of non-zero genes observed in each sample yielded consistent results as with our results from samples collected in experimental batch 2. Given the axon-TRAP data from the RGCs, the samples collected in the superior colliculus (SC) samples – the axonal RGC compartments-- have fewer non-zero counts compared to the samples collected from the retinal cell body (Figure 4.9). This indicates a compartment-specificity to the number of unique mRNA species observed.

These results are consistent with the non-zero genes observed in the ipRGC axon-TRAP study's experimental batch 2. In the experimental batch 2, I see more unique mRNA species represented in the translatome of neuronal cell bodies compared to that of axons, where I see fewer mRNA species which have specific axonal development/homeostatic functions. Samples taken from the retina should include all ribosome-bound mRNA from ipRGC cell bodies, performing a diverse set of

functional roles, while samples taken from the dLGN and SCN tissues are isolated ribosome-bound mRNA from the *axonal projections* of ipRGC. mRNA species packaged and localized to axonal compartments are likely to have axon-specific functional roles, and therefore should be a subset of all mRNAs translated in the cell.

If the experimental protocol for axon-TRAP-RiboTag are considered, the discrepancies between the number of non-zero genes quantified between experimental batch 1 and 2 begin to show their origins. In the axon-TRAP protocol, the goal is to isolate the cell-type specific ribosome-bound mRNA from the total pool mRNAs in the tissue. The efficiency of the immunoprecipitation step determines how specific the resulting pulldown is. If the IP is 100% effective, the resulting pulldown will be purely ribosome-bound mRNAs, i.e. the cell type-specific translatome, while if the IP is ineffective, the resulting pulldown will be the non-specific set of mRNAs in the tissue, i.e. the bulk transcriptome. To determine the efficiency of each experimental batch's translatome pulldown, I compared the number of non-zero genes in each sample of the axon-TRAP experiments to bulk-RNA sequencing samples collected in the same tissue types. The distinct differences between the number of non-zero genes observed between experimental batch 2 and the bulk transcriptome samples indicates the IP was at least in part successful in isolating the translatome (although the exact lower bound is not known) (Figure 4.8). Conversely, the experimental batch 1 samples show a trended progression from reflecting a transcriptomic-like number of non-zero genes observed in samples collected in 12/2019 to closer to a translatomic-like number of non-zero genes observed in

samples collected in 10/2020. This suggests that the isolation efficiency in samples collected in experimental batch 1 was variable and increased over the progression of the study, but never reaches that of experimental batch 2.

4.4 Conclusion

New, complex experimental protocols are pushing the bounds of attainable information explaining how cells develop, function, and maintain homeostasis at maturity. Optimization of these methods and congruent computational methods is necessary to ensure the validity of the biological implications drawn from them. Study-specific computational control metrics are extremely important when evaluating new datatypes to confirm the validity and consistency of the data collected. If data collection is inconsistent, variability may present as batch effect (unwanted/technical artifacts causing differences between samples) in datasets when in actuality it is a true reflection of biological diversity between samples *caused by* unintended differences in experimental techniques. In particular, in new experimental protocols isolating the translatome from the transcriptome, metrics such as the non-zero count are instrumental to diagnose the true biological and technical contributions in the gene expression heterogeneity.

4.5 Materials and Methods

All experimental data for this study was collected by Rashmi Gupta and Shyrice M. Mitchell. All data preprocessing and downstream analysis was performed by Theresa

A Alexander. Figures for this chapter were produced and edited by Theresa A Alexander.

4.5.1 Experimental Design

In order to evaluate changes to the local translatome in ipRGCs throughout development and in response to changes in the intrinsically-driven activity levels, neuronal tissue samples were collected in 3 different tissue regions—retina, suprachiasmatic nucleus (SCN), and the dorsal lateral geniculate nucleus (dLGN)—at 3 developmental ages—postnatal days 8, 15, and 60 (P8, P15, P60)—, and in 3 genotypes—Opn4+/-, Opn4-/-, and Opn4+/+. The samples were collected and mRNAs isolated by TRAP and processed in two batches of sequencing. The first set of samples were collected by experimentalist A and processed for library construction using an unknown library preparation kit and RNA-sequencing through a commercial company Genewiz. The second set of samples were collected by experimentalist B, processed for library construction using an ultra-low input library kit—TakaraBio SMARTer Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian Kit (Takara, 634486) with SMARTer RNA Unique Dual Index Kit – 96U (8bp) (384 reactions) (Takara, 634452)—and Illumina RNA-sequencing at the University of Maryland Brain & Behavior Institute - Advanced Genomic Technologies Core (BBI-ACGT). To ensure the presence of the HA-tagged ribosomal protein in the TRAP'ed pool of RNA, a western blot was conducted. After immunoprecipitation of the HA-tagged ribosome bound mRNA, we run the precipitate on a gel and label with HA antibody to visualize the presence of protein bands containing the HA epitope.

4.5.2 Preprocessing of RNA-Sequencing Reads

All sample reads were trimmed using either cutadapt (M. Martin, 2011) or trimmomatic (Bolger et al., 2014) (in accordance with the recommended pipeline for each library preparation kit) to trim sequencing adapters, unique molecular identifiers (UMIs) if the library preparation incorporated them, and any low-quality read sequences defined by a low base pair call score. After trimming each sequenced read, the result will be high quality biological sequence.

Trimmed reads were then mapped to the Genome Reference Consortium Mouse Build 38 (Schneider et al., 2016) using HISAT2 (T. Kim et al., 2019) and gene counts were obtained using HTSeq (Anders et al., 2015; Putri et al., 2022). All downstream analysis was completed using R version 4.2.1 (R Core Team, 2021) using custom functions developed by Theresa Alexander as well as functions developed by Ashton Trey Belew as part of the hpgltools package (https://github.com/elsayed-lab/hpgltools).

4.5.3 Cumulative F-Statistic with SVA Dimensions

Surrogate variable analysis (SVA) is a dimensionality reduction method which estimates surrogate variables (SVs) to model variation not attributed to the experimental condition of interest. SVA assumes that all unknown or unmodeled sources of variation are unwanted and is therefore commonly used as a batch correction method. Given a full model including the variable for the outcome of interest and a null model including only an intercept, SVA first identifies the number

of latent variables. Next, the SV weights are estimated, effectively modeling the variational sources that adequately model variational differences between the full and null model (i.e. all residual variation not modeled by the outcome variable of interest). In this case, the estimated SVs should model known sources of batch (experimental batch) well while modeling minimal variational contributions from the outcomes of interest.

After using SVA to estimate the SVs, I perform linear regressions which model each estimated dimension of SVA as a function of the observed outcome variable as well as known, wanted sources of variation (experimental batch, location, developmental age, and genotype).

$$\underbrace{X_i}_{\text{dimension i of SVA}} = \underbrace{B_0 + B_1 batch/location/age/genotype}_{\text{modeled underlying group structure}}$$

Then, to assess the degree to which each known source of variation, both wanted and unwanted, contributes to the estimated SVs, I use the F-statistic to measure the between-to-within class variance in each model, defined as:

$$F - statistic = \frac{TSS - RSS}{RSS}$$

where I compute the total sum of squares (TSS) and residual sum of squares (RSS) per regression model. This indicates how well the dimensional reductions for SVA describes group separation for known groups.

4.5.4 Variance Partition Analysis

Variance partition analysis evaluates the contribution of modeled sources of biological and technical variability both at a genome-wide summary scale, as well as on a per-gene basis. First, the variance explained by each variable in the input model is estimated by fitting a linear mixed model to quantify the contribution of each dimension of variation to each gene (Hoffman & Schadt, 2016). This can be reported globally across all genes or broken down to report at the gene-level. I employ variance partition analysis to evaluate the variational contributions of the modeled technical and biological variables of interest, as well as to identify genes whose variability is largely explained by the experimental batch variable.

4.5.5 Sequencing Depth Contributional Assessment

The number of genes quantified in sample i is defined as the non-zero gene count and is equal to the number of genes in sample i with counts greater than 0. This metric indicates the diversity in gene expression and the level of binary on-or-off expression globally across all genes.

To evaluate the relationship between sequencing depth and the non-zero gene count and the potential bias observed in the resulting genes observed in samples with low sequencing depth, I used a down sampling technique. In samples from experimental batch 1, where the average mapped read counts was roughly double the observed mapped read counts in experimental batch 2, each biological sample's counts were downsampled such that the total number of counts in the sample matched the library size of biological samples in experimental batch 2. In down sampling for each biological sample, the probability of sampling a count from gene i was equal to $\frac{observedcounts_i}{\sum_{n=1}^{k} observedcounts_n}$, resulting in matched library sizes in both experimental batches. I performed a bootstrapped down sampling for a total of 10 iterations.

Chapter 5: The role of melanopsin on ipRGC cell autonomous and cell non-autonomous programming

5.1 Introduction

5.1.1 Intrinsically photosensitive retinal ganglion cells

Within only the last few decades a third type of retinal photoreceptor has been characterized which modulates a wide array of biological processes, such as regulation of sleep/awake cycles, metabolism, hunger patterns and hormone production (Lazzerini Ospri et al., 2017). These retinal photoreceptors respond directly to light cues and have been shown to operate independently of rods and cones (D. M. Berson, 2002; Hattar, 2002). This new subclass of photosensitive neurons, the intrinsically photosensitive retinal ganglion cells (ipRGCs), project their axons from the ganglion cell layer of the retina to innervate several brain targets. The six different types are physiologically unique and are responsible for an array of light-dependent tasks. The diversity in function and morphology of these ipRGC types suggests that the transcriptional and/or translational landscape differs between them. One of the most abundant ipRGC types, M1, is exclusively responsible for the photoentrainment of the "master clock regulator", the Suprachiasmatic Nucleus (SCN) (Beier et al., 2020). The ipRGC M1 type cells are unique because they have a delayed developmental time course compared to conventional RGCs and begin to innervate the SCN postnatally (McNeill et al., 2011). The specialization of ipRGC types based on their innervation of distinct central targets could be accomplished by differentially shuttling mRNA transcripts to axonal terminals where monosomes and polysomes transcribe them to proteins (Holt et al.,

2019). The local protein translation profile would therefore be differentiated based on developmental stage and cell type (Hafner et al., 2019; K. C. Martin, 2004; Pfeiffer & Huber, 2006). Together, this indicates that the development and functional needs of soma versus axonal processes of ipRGCs are distinct and may be coordinated by transcriptional and translational regulatory shifts.

5.1.2 Melanopsin and its role in development

ipRGC axon development and early synaptogenesis in central targets occurs during a period where the photosensitivity of the retina is determined solely by the presence of the photosensitive protein, melanopsin (Opn4) (D. M. Berson, 2002; Fu et al., 2005; Hattar, 2002; Provencio et al., 1998; Renna et al., 2011). Dysregulation of Opn4 has shown to have vast consequences within the developing retina and innervation targets.

One such consequence is on the spontaneous activity in developing retinas known as retinal waves. Retinal wave properties are regulated in part by light through ipRGCs (Kirkby & Feller, 2013). When cholinergic circuits were blocked, the drivers of retinal waves in early post-natal development, compensatory retinal waves were generated through ambient light-activated signaling of ipRGCs (Kirkby & Feller, 2013). Within the retina, ipRGCs influence retinal waves by a gap-junction coupled network with other retinal neurons (Arroyo et al., 2016; Pérez de Sevilla Müller et al., 2010; D.-Q. Zhang et al., 2008), rendering the retina light-sensitive before canonical photoreceptor maturity (Sernagor, 2005). Previous studies have shown that when Opn4 is knocked out, light no longer affects wave burst duration as it does with Opn4 expression,

suggesting ipRGCs are critical players in the link between photic input and retinal wave properties (Renna et al., 2011).

Knocking out Opn4 also has morphological consequences within the retina and brain innervation targets in addition to the physiologic changes. Previous studies have shown that mice that are either dark-reared or have Opn4 knocked out have inappropriate overgrowth of retinal vasculature because of a dysregulated hypoxic environment leading to reduced vascular endothelial growth factor (VEGFA) expression (Rao et al., 2013). Additionally, the light-signaling pathway through ipRGCs is critical to maintain not only retinal wave properties but also retinofugal projection development (Renna et al., 2011; Tiriac et al., 2018). Specifically, light deprivation inhibiting melanopsin-dependent activity within ipRGCs inhibits eye-specific segregation of retinohypothalamic (RHT) projections as well as retinogeniculate projections (Renna et al., 2011; Tiriac et al., 2018).

Though several functional differences have been identified resulting from lack of melanopsin expression, the direct transcriptomic and translatomic shifts caused by knocking out melanopsin in ipRGCs have not been characterized. Understanding the molecular determinants within ipRGCs and their connectome will help elucidate pathways which modulate these physiological and morphological manifestations directed by Opn4. Activity-dependent local protein production is a regulatory mechanism which may shed light on regulatory shifts within ipRGCs which drive the downstream dysregulation. First, subcellular localization of mRNAs occurs by cis-

acting elements such as RNA binding proteins (RBPs) and microRNAs silencing mRNAs that in turn get shuttled to axonal compartments in formed ribonucleoprotein (RNP) complexes (Andreassi et al., 2018). This localized, translationally suppressed pool of axonal mRNAs can then be translated to protein "on-demand" to support axonal growth, axon guidance, injury responses, or to sustain long term potentiation/depression (LTP/LDP) when a stimulus prompts the process of spatially controlled protein synthesis (Jung et al., 2012; Li et al., 2021). Stimuli which initiate local protein translation can be external factors such as growth cues like Netrin-1 (Leung et al., 2006; Yao et al., 2006) or internal cues such as neuronal activity (Choe & Cho, 2020). In particular, increased neuronal activity leads to an increase in local translation mediated by microRNAs (miRNAs) and RNA binding proteins (RBPs) in the neuropil (Bartley et al., 2016; Narayanan et al., 2007; Sutton et al., 2004). Because of this evidence of the regulation of local translation by neuronal activity, melanopsin expression dictating action potentials evoked by photic stimuli may play an important role in regulating activity-dependent plasticity of local protein production in the development of ipRGC circuits.

5.1.3 Experimental Design

Though the efforts to measure transcript populations and functions in a diverse set of neuronal cell types have been great, the exact mRNA populations in local subcompartments of ipRGCs and their transcriptional and translational regulation in response to changes in melanopsin-driven activity have yet to be characterized. Elucidating the specific changes to global transcriptomes as well as local translatomes

in the presence and absence of Opn4 will give insight into both ipRGC intrinsic translatomic changes and extrinsic transcriptomic changes. In order to understand the role melanopsin plays in ipRGC and retinal development and function, we used bulk RNA-sequencing (RNA-seq) as well as axon-TRAP-RiboTag to measure ipRGC-extrinsic changes in gene expression as well as ipRGC-intrinsic changes in translation in melanopsin knockout mice.

The bulk transcriptome was measured at P8 (before eye-opening) and at three different regions of the ipRGC connectome: the retina where ipRGC somas originate, the dLGN where non-M1 types of ipRGCs project to, and the SCN, where a majority of M1 ipRGC types innervate. These samples were collected in both melanopsin (Opn4) present (Opn4 "Hets") and melanopsin absent (Opn4 "KOs") to evaluate the global transcriptional shifts in response to a loss of melanopsin-driven activity.

In addition, the ipRGC-specific translatome was measured at P15 (after eye-opening) at the same three regions to evaluate the cell type-specific sub-compartmental translatomes. These samples were also collected in Opn4 "Hets" and Opn4 "KOs" to evaluate the cell type-specific sub-compartmental shifts in translation with the presence and absence of melanopsin-driven activity.

5.2 Results

5.2.1 Gene expression is tissue location and Opn4 genotype specific

I was interested in comparing the global, tissue-wide transcriptomic signatures between the retina, SCN, and dLGN in mice that express Opn4 and mice that lack Opn4. In order to compare tissue-level gene expression changes between animals with one functional copy of melanopsin (Opn4 Cre/+) and melanopsin knock-out (KO) (Opn4 Cre/Cre) animals, samples collected at postnatal day 8 (P8) were subjected to a bulk transcriptomic analysis. Principal component analysis (PCA) of the counts per million (CPM) normalized counts for samples that passed the filtering process (n = 15) reveals a significant variational contribution (46.35% variance) in expression that is caused by sample location, which distinguishes samples from the retina (green) from those taken from brain target regions (dLGN - pink, SCN - blue) (Figure 5.1A). Secondarily, genotypic global differences across all locations were detected in the PC2 projection accounting for 31.86% variance (circles = Opn4 hets, triangles = Opn4 KO). A heatmap of the z-score normalized CPM counts for the top 200 most variable genes shows the same consistent clustering both by sample location (pink, blue, green top bar) as well as sample genotype within each location (color tone changes top bar) (Figure 5.1B).

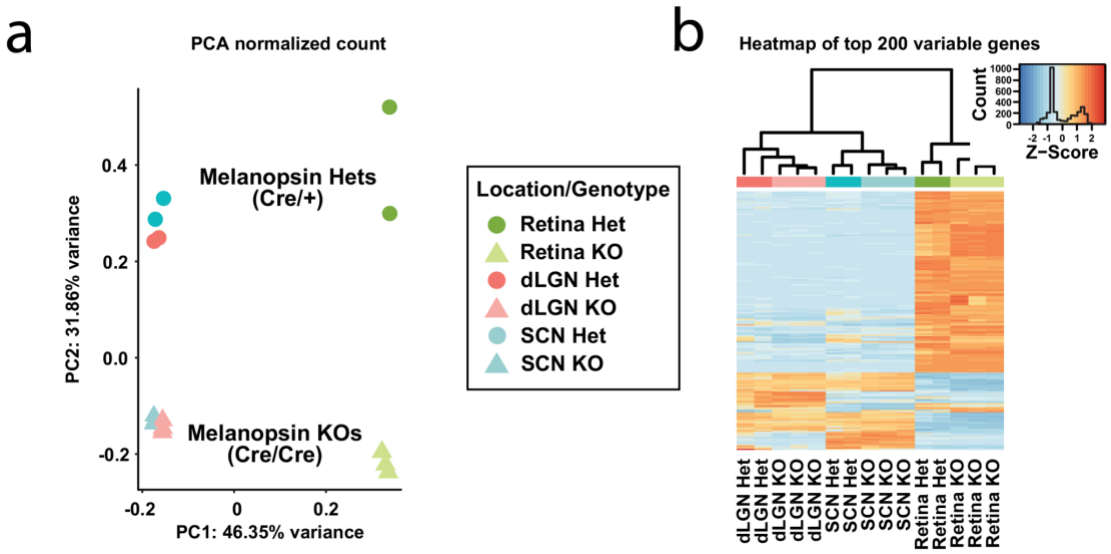


Figure 5.1. ipRGC subcompartment and Opn4 (melanopsin) genotypic differences in global transcriptomics.
A) The PCA of the log2 CPM normalized counts reveals clustering specific to genotype in the PC2 projection and retina versus brain target clustering in the PC1 projection. B) A heatmap of the top 200 variable genes' Z-score normalized expression across all samples reveals distinct genotype- and location-based clustering.

Differential gene expression (DE) analysis comparing pairwise contrasts of Opn4 genotypes within each location at P8 revealed global dysregulation (Figure 5.2A). The exact number of DE genes found to be significantly dysregulated at each location between Opn4 genotypes can be found in

The retina transcriptome had a total of 9518 DE genes, the SCN had a total of 5154 DE genes, and the dLGN had a total of 5985 DE genes (log2FC >= 1, adjusted P value < 0.05) between the Opn4 genotypes (Figure 5.2B, Table 5.1).

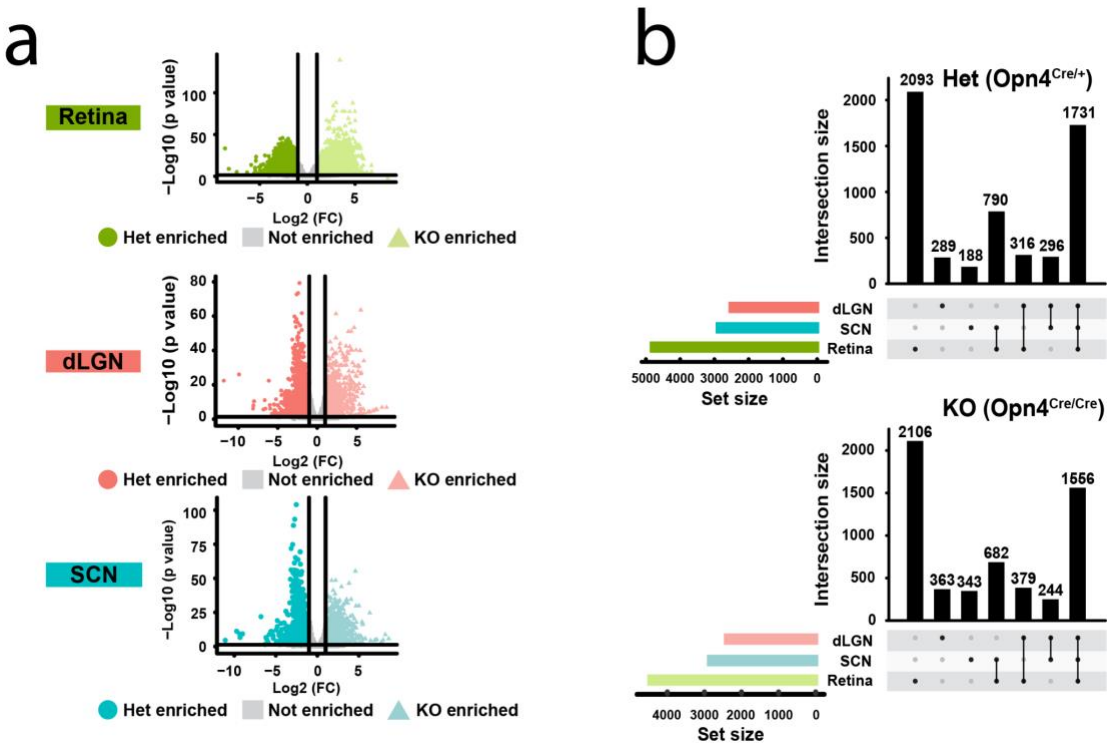


Figure 5.2. Differential gene expression (DE) analysis of the transcriptome.
A) Pairwise DE comparisons within each location show dysregulation between Opn4 genotypes. B) Upset plot for the set of DE genes dysregulated in Opn4 hets (top) and Opn4 KOs (bottom) show both shared and unique dysregulation of mRNA expression between locations.

I was interested to compare the transcriptomic shifts between tissue locations within each Opn4 genotype. The significantly upregulated DE genes (log2FC > 1) in each location were subsetted and then examined for shared/distinct signatures (Figure 5.2B, top). In the Opn4 Hets, I found the retina had the most unique DE gene set based when compared to the number of unique genes upregulated in the dLGN or SCN (2,093 genes uniquely upregulated in the retina Opn4 hets, 289 genes uniquely upregulated in the dLGN Opn4 Hets, and 188 genes uniquely upregulated in the SCN

Opn4 Hets). Transcriptomes from all three locations share 1,731 genes which are upregulated in Opn4 Hets compared to KOs.

Opn4 KO dysregulation has both shared and unique expression signatures between locations. When comparing the transcriptomic profiles of genes upregulated in Opn4 KOs to Hets of the same location, 2106, 363, and 343 genes were specifically upregulated in the retina, dLGN, and SCN, respectively. Additionally, 1,556 genes were upregulated in Opn4 KOs across all locations.

Table 5.1. Significant DE genes dysregulated between Opn4 genotypes.

	Opn4 Het Enriched	Opn4 KO Enriched
Retina	4,930	4,588
dLGN	2,632	2,522
SCN	3,005	2,980

Within the top significant DE genes dysregulated in the retina were genes belonging to the PDGF family–receptors PDGFRA and PDGFRB, and ligands PDGFB, PDGFD) (Table 5.2). Receptor PDGFRA and ligand PDGFD were found to be upregulated in Opn4 Hets, while receptor PDGFRB and ligand PDGFB were found to be upregulated in Opn4 KOs. This suggests perturbation of melanopsin-driven activity contributes to downstream effects on PDGF signaling involving these receptor-ligand pairs.

Table 5.2. DE results for receptor/ligand pairs in the PDGF family in the retina.

	Log_2 Fold Change	Adjusted *P* Value
PDGFRA	2.149 (Opn4 KO Enriched)	8.991e-19
PDGFRB	-2.238 (Opn4 Het Enriched)	1.223e-14
PDGFB	-3.07 (Opn3 Het Enriched)	1.313e-20
PDGFD	2.351 (Opn4 KO Enriched)	5.717e-8

Ephrin receptors and ligands are critical signaling factors important for sprouting angiogenesis and vascular morphogenesis (Adams et al., 1999). Several ephrin receptors (Eph) and ligands (Efn) were differentially regulated in the retina transcriptome when comparing Opn4 Hets versus Opn4 KOs (Table 5.3). Both type A and B ephrin ligands were found to be significantly differentially expressed, with Efn A1, A2, A3, and B1, B3, and B4 enriched in Opn4 Hets and just Efn A5 and B2 upregulated in Opn4 KOs. Of the corresponding ephrin receptors, all receptors significantly DE between Opn4 Hets and Opn4 KOs—Eph A1, A8, B4, and B6—were upregulated in Opn4 Hets.

Table 5.3. DE results for receptor/ligand pairs in the Ephrin family in the retina.

Significantly Dysregulated Ephrin Ligands		
Gene	Log2FC	Adj P Value
EfnA1	-2.529 (Opn4 Het Enriched)	1.549e-13
EfnA3	-2.24 (Opn4 Het Enriched)	3.032e-17
EfnB3	-1.813 (Opn4 Het Enriched)	1.093e-20
EfnB1	-1.759 (Opn4 Het Enriched)	1.662e-14
EfnB4	-1.643 (Opn4 Het Enriched)	8.442e-8
EfnA2	-1.466 (Opn4 Het Enriched)	7.665e-17
EfnB2	1.312 (Opn4 KO Enriched)	0.00002454
EfnA5	1.362 (Opn4 KO Enriched)	0.0000141

Significantly Dysregulated Ephrin Receptors		
Gene	Log2FC	Adj P Value
EphA8	-2.979 (Opn4 Het Enriched)	8.965e-18
EphA1	-2.896 (Opn4 Het Enriched)	0.00004375
EphB4	-2.895 (Opn4 Het Enriched)	9.116e-17
EphB6	-2.402 (Opn4 Het Enriched)	3.798e-17

To investigate the functional properties of the transcriptomic dysregulated between Opn4 hets and Opn4 KOs within each location, I conducted gene set enrichment analysis (GSEA) using several gene annotation databases including Gene Ontology (GO), KEGG, and REACTOME. Beginning with the list of significantly enriched DE genes from each location's Opn4 Het versus Opn4 KO contrast, GSEA identified several significantly enriched location-specific functional pathways (Figure 5.3). The top significant hits from GO cellular compartment, molecular function, and biological processes categories include several receptor activity pathways (ephrin receptor activity, retina; norepinephrine binding, dLGN) as well as indicators of shifts in ribosomal machinery within the brain compartments (dLGN and SCN).

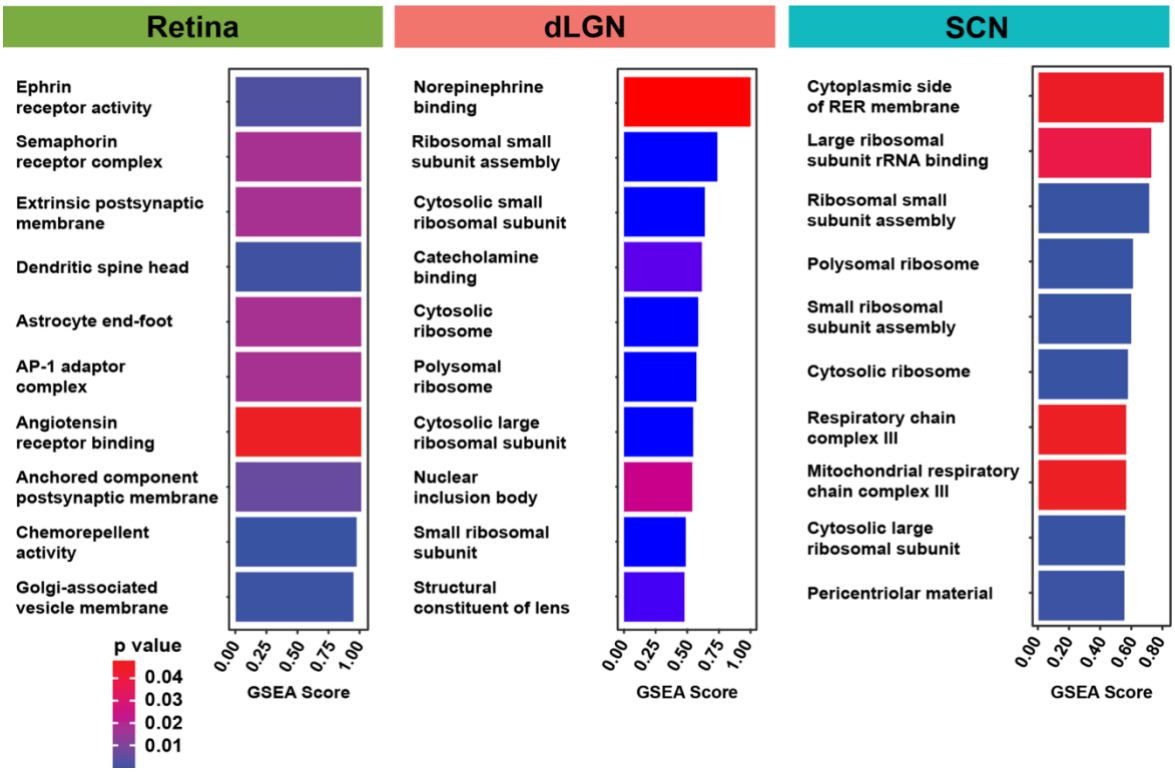


Figure 5.3. GSEA of ipRGC transcriptome.
Top significant hits from gene set enrichment analysis (GSEA) of transcriptomic expression shifts between Opn4 hets and Opn4 KOs in the retina (left, green), the dLGN (middle, pink), and the SCN (right, blue) include several receptor activity pathways as well as indicators of shifts in ribosomal machinery.

5.2.2 Local translation occurs in ipRGC axons at postnatal day 15

I wanted to evaluate not only the global, tissue-wide shifts in the transcriptome associated with a loss of Opn4, but also the cell-autonomous translatomic shifts within the Opn4-expressing ipRGCs. Using a knock-in mouse that expresses Cre-recombinase in place of endogenous melanopsin ($Opn4^{Cre}$), we labeled ribosomal proteins with a floxed $Rpl22^{HA}$ line and developed an ipRGC-specific Translating Ribosome Affinity Purification (TRAP) assay to isolate the ipRGC-specific translatome. To examine pathway-specific local protein translation, we

immunopreciptated HA-tagged ribosomes, isolated associated translating mRNA and sequenced the associated translating mRNAs from ipRGC cell bodies in the retina and axons in the dorsal lateral geniculate nucleus (dLGN) and suprachiasmatic nucleus (SCN). We performed these experiments after eye-opening (P15) in both control ($Opn4^{+/Cre}$::$Rpl22^{HA/+}$) and melanopsin knockout mice ($Opn4^{Cre/Cre}$::$Rpl22^{HA/+}$).

To determine if there were differences in the number of mRNA species translated in ipRGC subcompartments, I evaluated the number of nonzero genes observed within each sample (as described in Chapter 4). The number of nonzero genes observed is differentiated by retinal compartment (translatome of ipRGC cell bodies, pink) versus brain target compartments (translatome of ipRGC axonal compartments, green and blue) (Figure 5.4). Conversely, there were no differences in the number of nonzero genes observed due to genotype of the samples (shape).

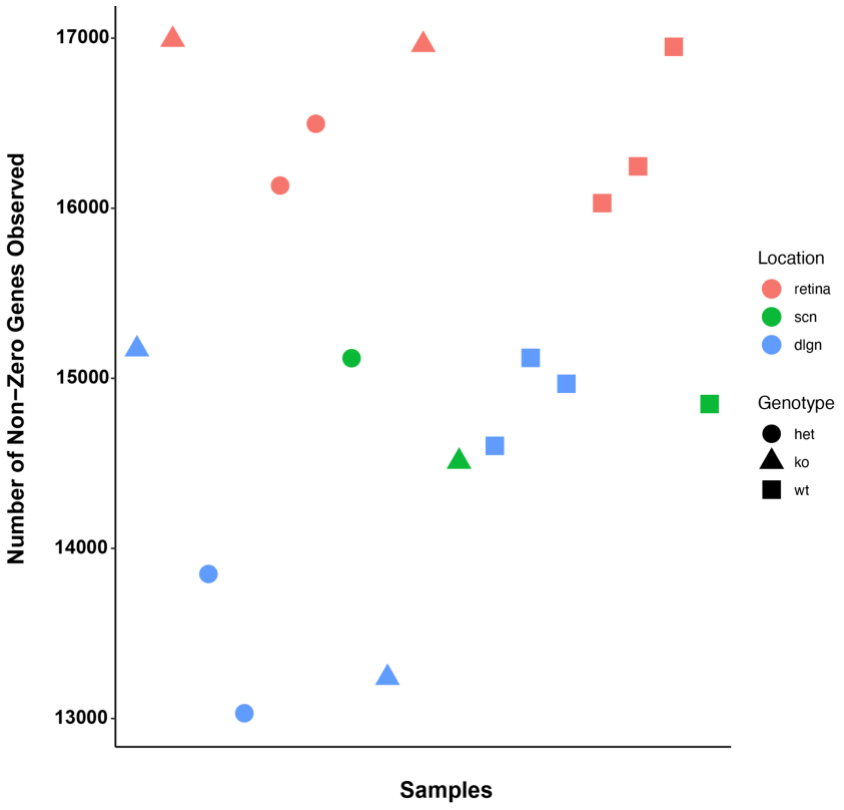


Figure 5.4. Non-zero genes observed in study samples.
The number of unique mRNA species (defined as "non-zero genes observed") expressed in each sample. Retinal samples (ipRGC cell bodies, pink) show an increased number of unique genes observed, while samples from the dLGN (blue) and SCN (green) (ipRGC axonal subcompartments) have fewer mRNA species translated.

Conventional retinal ganglion cells (RGCs) have been demonstrated in earlier studies to transport mRNAs to axonal compartments where local translation occurs (Shigeoka et al., 2016). I sought to determine whether local translation also occurs in ipRGCs and, if so, which mRNA species are preferentially localized and translated in ipRGC axonal compartments and which are translated in ipRGC soma bodies in the retina. I conducted a DE analysis comparing the ipRGC somal subcompartment to the axonal subcompartments projecting to the dLGN at postnatal day 15 (P15). I identified 2,638

mRNA species upregulated in the ipRGC somal compartment in the retina (pink) and 2,405 mRNA species upregulated in the ipRGC axonal compartments projecting to the dLGN (green) (Figure 5.5A). This suggests that there is preferential localization of mRNAs within ipRGCs projecting to the dLGN and local protein translation occurs within these ipRGC subcompartments.

To discover the functional classes these sub-compartment enriched translated mRNAs are part of, I calculated a gene set enrichment score for GO terms associated with "retina" and retinal function as well as "axon" and axonogenesis related terms. Consistent with local translation signatures from RGCs (Shigeoka et al., 2016), when compared to the dLGN axonal compartments, ipRGC translatomic signatures related to retinal function are upregulated in the retina. ipRGC translatomic signatures enriched in dLGN axonal compartments generally encode proteins with functions associated with axonal development (Figure 5.5B). This suggests that axonogenesis related functions for ipRGC axonal subcompartments projecting to the dLGN may be mediated by local mRNA translation.

After confirming signatures of axonal protein synthesis is the dLGN ipRGC subcompartment, I wanted to identify the other enriched functional groups in an unsupervised manner. I performed GSEA of the mRNAs upregulated in the dLGN axonal subcompartment. A large signature of mitochondria-associated translation at P15 was identified as nine out of the top ten significant GSEA hits with GSEA recall score > .6 have functions associated with ATP synthesis and mitochondria support

and function (Figure 5.6). This suggests at P15, local protein translation of mitochondrial proteins aids in maintaining mitochondrial health so that ATP energy requirements can be met during this developmental stage.

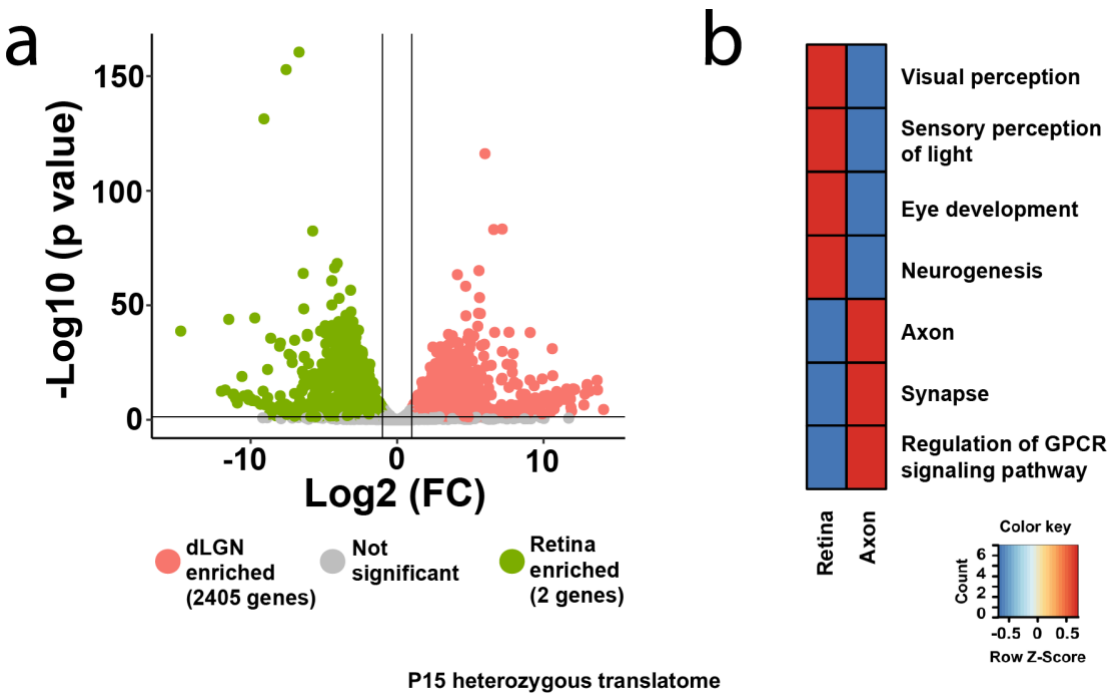


Figure 5.5. ipRGC subcompartment-specific mRNA translation.
A) DE results between the ipRGC dLGN and retina subcompartmental translatomes. The dLGN (green) has 2,405 genes with upregulated translation while the retina (pink) has 2,638 with upregulated translation. B) Average expression of "soma" and "axon" related gene ontology (GO) pathways in retina and dLGN samples.

To determine the functional groups of genes which are translated locally in dLGN-projecting ipRGC subcompartments, I conducted GSEA using the set of genes upregulated in the dLGN subcompartment compared to the retina (2405 genes, Figure 5.5). GSEA of these genes revealed a large translational signature of genes important

for supporting mitochondrial health and energy production (Figure 5.6). This suggests that mitochondria are supported and maintained by local protein translation within axonal subcompartments during synaptogenesis.

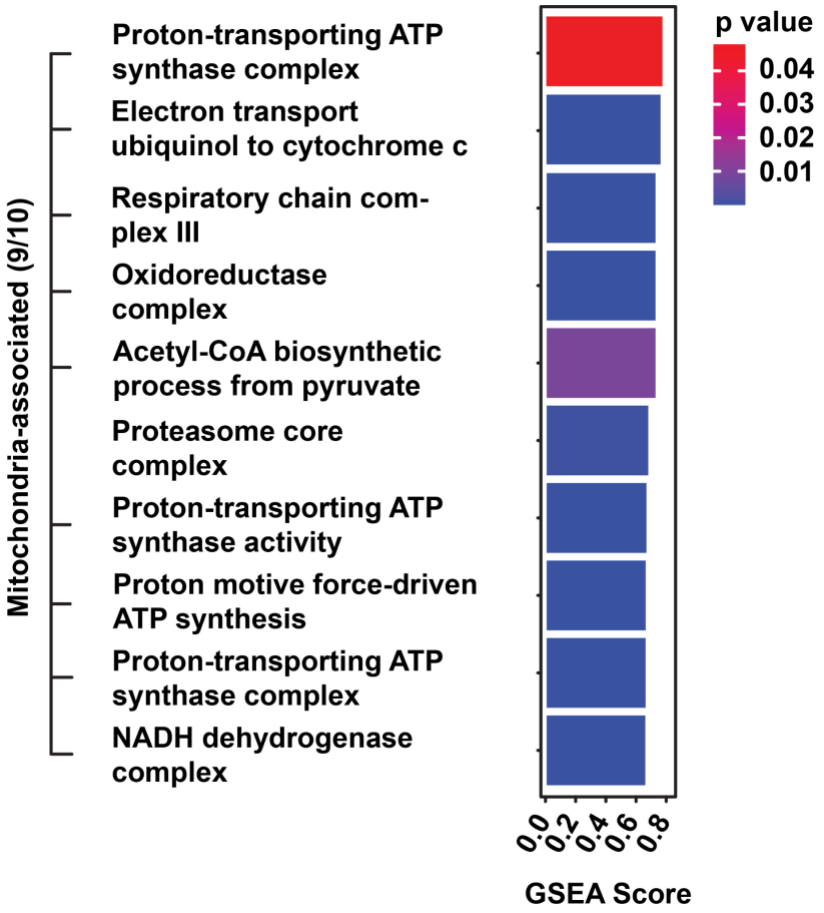


Figure 5.6. GSEA of axon-enriched mRNAs.
Gene set enrichment analysis (GSEA) of genes which are enriched in ipRGC axonal subcompartments projecting to the dLGN. Nine out of the top ten top GSEA significant hits have functional annotations related to sustaining mitochondrial health.

5.2.3 Local translation in ipRGC somas is melanopsin-dependent at P15

To identify if local translation within retinal ipRGC subcompartments is dependent on melanopsin expression, we performed a preliminary DE analysis (pending more samples being collected) comparing retinal translatome samples with (Opn4 Hets, n = 2) and without (Opn4 KOs, n = 2) samples (Figure 5.7). A total of 309 genes were found to have dysregulated translation in the retinal ipRGC subcompartment, with 256 genes upregulated in Opn4 Hets and 53 genes upregulated in the Opn4 KOs. As confirmation, Opn4 translation has the largest fold change difference between Hets (where it is expressed) and KOs (where it has been experimentally knocked-out). This suggests the absence of Opn4 expression in ipRGCs causes a shift in translation within ipRGC somas in the retina at P15.

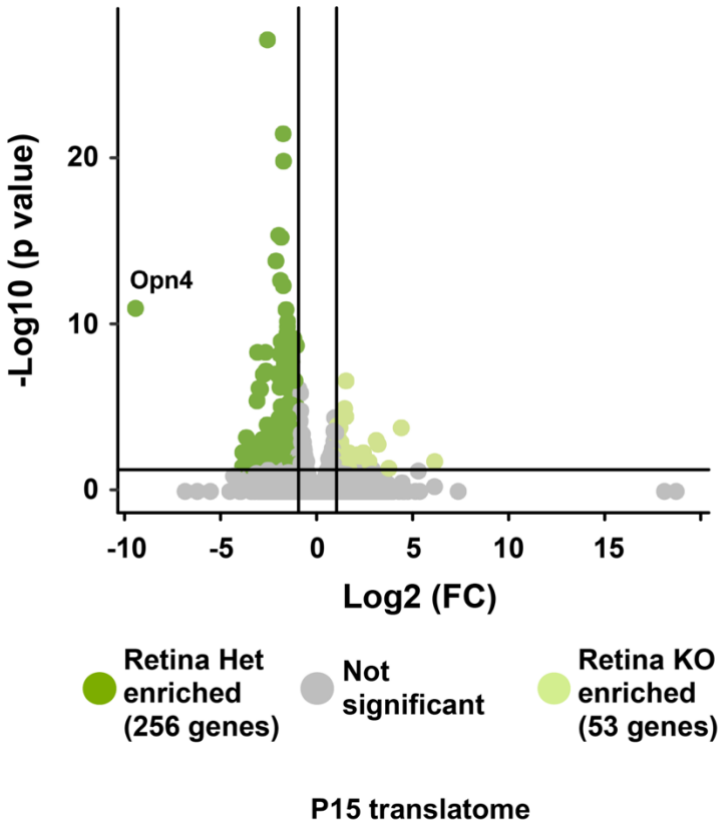


Figure 5.7. Retina Opn4-dependent translation.
Comparison of ipRGC translatomes in the retina (ipRGC somas) in the presence (n = 2) and absence (n = 2) of Opn4. A total of 309 genes were found to be dysregulated between retina Opn4 Hets and retina Opn4 KOs with an absolute value of the log2FC cutoff > 1 and FDR adjusted *P* value cutoff of 0.05.

5.2.4 Local translation in ipRGC axonal projections in the dLGN is melanopsin-independent at P15

After showing protein translation occurs locally in ipRGC axonal subcompartments and that there are shifts in translation between Opn4 genotypes within ipRGC somas, I aimed to determine if these translatomic shifts in the absence of Opn4 also occurred in the axonal subcompartment. I performed a DE contrast between ipRGC translatomic samples collected in the dLGN axonal subcompartments between Opn4

Hets and Opn4 KOs (Figure 5.8). The translatome within the axonal subcompartments of ipRGCs projecting to the dLGN remains essentially unaffected in the absence of Opn4 expression. Just 3 genes show a modest shift in translational regulation, Pianp, Sdhaf3, and Vipr2, which are enriched in the Opn4 KOs. This suggests that the P15 translation occurring in ipRGC axonal subcompartments in the dLGN is melanopsin-independent, but this will need to be verified by increasing the statistical power with the addition of more samples.

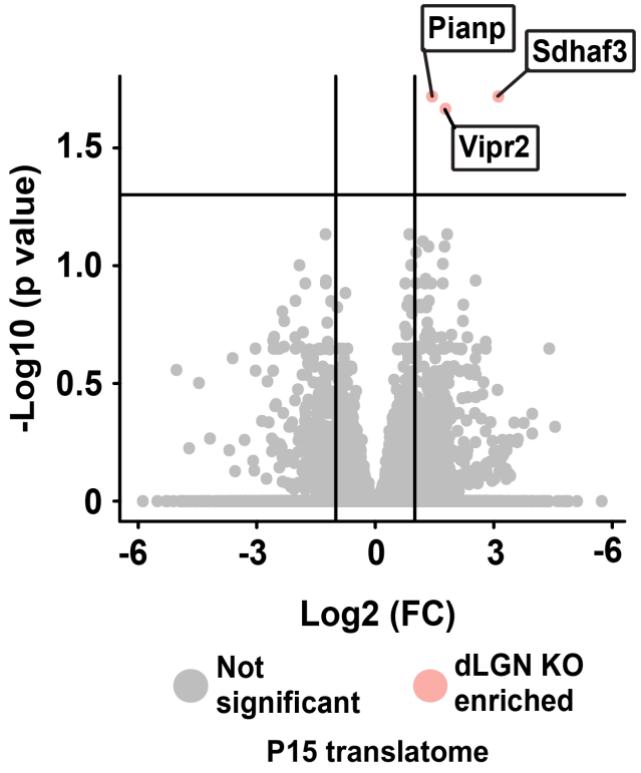


Figure 5.8. dLGN Opn4-independent translation.
Differential expression (DE) comparing the P15 translatomes of ipRGC axonal subcompartments projecting to the dLGN with (Hets) and without (KOs) Opn4 expression shows just 3 genes which have dysregulated translation.

5.3 Discussion

In this study, I tested the hypothesis that melanopsin signaling regulates both the global transcriptome in the retina and ipRGC projecting brain targets as well as the intra-ipRGC local protein translation within cell subcompartments. Using a knock-in mouse that expresses Cre-recombinase inserted in exon 4 of endogenous melanopsin ($Opn4^{Cre}$), we used two collection protocols to isolate the tissue-wide transcriptome for bulk RNA-sequencing as well as isolating labeled ribosomal proteins from a floxed $Rpl22^{HA}$ line using a novel ipRGC-specific Translating Ribosome Affinity Purification (TRAP) assay. I found 1) the global transcriptome was broadly dysregulated in the absence of Opn4 expression in the retina, SCN, and dLGN in early development (P8), 2) ipRGCs participate in local protein translation during axonogenesis and synaptogenesis at P15, and 3) ipRGC local protein translation is melanopsin-dependent within the retinal subcompartment but is melanopsin-independent within the dLGN-projecting axonal subcompartments at P15.

To identify global transcriptomic shifts, I first evaluated the bulk transcriptome at P8 for three tissue types (retina, SCN, and dLGN) in both control and melanopsin knockout mice ($Opn4^{Cre/Cre}$::$Rpl22^{HA/+}$). Principal component analysis revealed global differences in the transcriptomes of tissues part of the ipRGC connectome (retina, where ipRGC somata/dendrites reside, as well as axonal projections in the SCN and dLGN) (Figure 5.1). Within each tissue location, I observed the transcriptome was dysregulated on a global scale when there was no functional Opn4 within ipRGCs.

The transcriptomic shifts within the retina between Opn4 Hets and Opn4 KOs identified several critical signaling pathways including VEGF, PDGF and Eph signaling pathways. Previous studies have reported signaling pathways implicated in retinal vascular diseases and angiogenesis including vascular endothelial growth factor (VEGF) signaling, platelet-derived growth factor (PDGF) signaling, and ephrin (Eph) signaling (Afzal et al., 2007; L. A. Kim & D'Amore, 2012; O'Neal et al., 2013; Rao et al., 2013). Hypoxia has been well described as a regulatory feature of angiogenesis and hypoxia-induced factor 1a (HIF-1a) has been shown to have correlated expression with PDGF, VEGF, and Eph family genes (Clara et al., 2014; Krock et al., 2011; Vihanto et al., 2005). These signaling pathways all contain receptors classified as receptor tyrosine kinases (RTKs) which act as signal transducers critical in regulating angiogenesis (Lemmon & Schlessinger, 2010; Sudhesh Dev et al., 2021). Of these pathways, VEGF signaling at embryonic day 16 has been the only pathway explored as part of the Opn4 KO vascular defects to date (Rao et al., 2013).

Because HIF-1a was found to be significantly downregulated in Opn4 Hets when compared to Opn4 KOs in the retina (log2FC = 3.553, adjusted *P* value = 1.773e-17), I would expect to find downstream hypoxia-correlated signaling pathways to also be perturbed. Although I do not find the same factor, VEGFA, (potentially because of differences in the mouse line used in these experiments) found in Rao et al. to be directly differentially expressed in the retina transcriptome, several top DE genes have been described in other Hif-1a correlated signaling pathways: PDGF and Eph.

The role of platelet-derived growth factor (PDGF) signaling during retinal development and vascularization (Andrae et al., 2008; Sadiq et al., 2016). Inhibition of PDGF causes a reduction in corneal vascular density (Dell et al., 2006). Two PDGF receptors (PDGFRA, PDGFRB) and two PDGF (PDGFB, PDGFD) ligands were found to be significantly dysregulated in the retina between samples which have Opn4 expression and those that do not (Table 5.2). In PDGF ligand-receptor interactions, PDGF ligands first form a dimer pair before an interaction with the PDGF receptor occurs. PDGF-DD ligand pairs only have high affinity for PDGFRB receptors, while PDGF-BB ligand pairs will bind both PDGFRA and PDGFRB receptors (Donovan et al., 2013; Mamer et al., 2017). The interactions in the PDGF then drive connection, maturation, and remodeling of blood vessels. The dysregulation of the PDGF signaling pathway between Opn4 hets and Opn4 KOs suggests melanopsin-driven activity in ipRGCs play an upstream role in PDGF signaling, and perhaps dysregulation of PDGF family ligands and receptors contribute to vascular developmental defects. And more, it was recently discovered VEGFA ligand has a high affinity for not only VEGF receptors but also both PDGFRA and PDGFRB receptors (Mamer et al., 2017) which corroborates that both VEGF and PDGF signaling contribute to melanopsin-activity driven retinal vascularization.

Ephrin (Eph) signaling is another pathway which has been shown to coordinate retinal vascularization (Davies et al., 2010; Kaczmarek et al., 2021). At P8, the retinal expression of ephrin receptors and ligands is perturbed without Opn4 expression

(Table 5.3). The dysregulation of the PDGF and ephrin pathways may be evidence to support that the lack of melanopsin drives dysregulation of several RTK receptor families impacting retinal vascular development.

The global transcriptomes of the SCN and dLGN were also perturbed in response to Opn4 expression changes at P8. Based on a DE analysis comparing Opn4 Hets to Opn4 KOs, global expression shifts in the dLGN and SCN were observed (Figure 5.2A). GSEA reveals that the expression changes within each location are associated with changes in ribosomal machinery in each tissue type (Figure 5.2B).

To examine the intra-ipRGC regulatory changes impacting the tissue-wide transcriptomic shifts in response to lack of Opn4 expression, we isolated ribosome bound mRNA and sequenced the associated translating mRNAs from ipRGC cell bodies in the retina in Opn4 Hets and Opn4 KOs. I determined ipRGCs do participate in subcompartment-specific local protein translation in both the somal and dLGN-projecting axonal subcompartments in mice with normal Opn4 expression (Figure 5.5). Interestingly, I find genes preferentially trafficked and translated in the dLGN-projecting ipRGC axonal subcompartments have a major part in supporting local mitochondrial health and homeostasis (Figure 5.6). During development, axonogenesis and synaptogenesis are highly energy demanding. Previous studies have shown mitochondria ATP production to support the assembly of synapses during presynaptic development (Lee & Peng, 2008) as well as support energy requirements (Harris et al., 2012). Our results suggest that with unperturbed melanopsin-driven

activity, critical mitochondrial health-supporting proteins are translated locally in dLGN-projecting ipRGC axonal subcompartments.

I observed compartment-specific dependency when considering the translational shifts within ipRGCs as a consequence of knocking out Opn4. Within the retinal ipRGC sub-compartments, over 300 mRNA species were found to have altered translation in response to a lack of melanopsin-driven activity within ipRGC somas (Figure 5.7). ipRGC gene markers part of the phototransduction cascade are particularly susceptible to translational shifts in response to a lack of melanopsin-driven activity. Previous studies have described several novel ipRGC markers—genes with enriched expression in ipRGCs compared to conventional RGCs (cRGCs)—which are indicated in the Opn4 phototransduction cascade including diacylglycerol (DAG) signaling factors Rasgrp1 and Dgkg (Berg et al., 2019). I find both of these genes to be dysregulated in the ipRGC retinal translatome and enriched in Opn4 Hets compared to Opn4 KOs (Rasgrp1 log2FC = 1.004, adj *P* value = .037; Dgkg log2FC = 1.389, adj *P* value .003). This suggests that Rasgrp1 and Dgkg are involved in the ipRGC phototransduction cascade and when the cascade is disrupted by knocking out Opn4 in ipRGCs, translation of these genes is reduced. This may be because of a reduction in translation due to the lack of melanopsin-driven activity but could also be a result of less transcription of these genes, resulting in fewer available mRNA transcripts to be translated.

Interestingly, although the local translatome is melanopsin-dependent within the retinal ipRGC subcompartments, this is not the case for the dLGN-projecting axonal compartments at P15 (Figure 5.8). There are only three genes which are reported to be significantly differentially translated in the dLGN ipRGC subcompartments (Figure 5.8) —Pianp, Sdhaf3, Vipr2. This suggests that the global transcriptomic shifts seen in the dLGN is not a result of a shift in local protein translation with the innervating ipRGCs. Given the number of samples used for this analysis, further analysis which is fully powered is necessary.

5.4 Conclusions

Intrinsically photosensitive retinal ganglion cells (ipRGCs) are essential in early development to relay photic information from the outside world to the developing retina and brain. Deletion of the photopigment melanopsin (Opn4) perturbs melanopsin-driven activity within ipRGCs and has drastic developmental effects in the retina and brain innervation targets. I identified major global gene expression changes across the tissues that contain ipRGC subcompartments including the retina, SCN, and dLGN when Opn4 is dysregulated. The transcriptomic shifts in the retina include several signaling pathways important for angiogenesis including PDGF and Ephrin signaling pathways, indicating that perturbation of melanopsin-driven activity in ipRGCs may cause dysregulation of angiogenesis through these pathways.

Additionally, I discovered mRNAs are preferentially shuttled to axonal compartments and local protein translation occurs in the subset of ipRGCs projecting to the dLGN at P15. Although global transcriptional shifts were observed in the dLGN transcriptome when Opn4 is knocked-out at P8, this is may not be due to a shift in local protein translation in dLGN-projecting ipRGC axonal subcompartments (although these axon-TRAP results are from a different developmental time point, P15).

5.5 Future Directions

This study has brought intriguing insights into the impact of Opn4 expression and Opn4-driven activity on developmental shifts in tissues part of the ipRGC connectome. To complete this study, a full complement of samples needs to be sequenced to ensure adequate statistical power for these analyses, particularly for the translatomic portion of the study. Additionally, I would like to look further into the Opn4-cre mouse line. I observed some inconsistencies with Opn4 expression in mice with the genotype Opn4 cre/+ (Opn4 Hets) (Supplementary Figure 5.1). I determined that these expression inconsistencies were not due to polymorphisms within the Opn4 coding region or an imbalance in Opn4 isoform expression (Supplementary Figure 5.2).

To determine if there was genetic drift of the Opn4-CRE mouse line, I used the variant caller Freebayes to determine positions of variants within each bulk RNA-seq retina sample (3 Opn4 Hets nad 3 Opn4 KOs) (Garrison & Marth, 2012). The jaccard index was computed between each pairwise set of SNPs called for each sample

(Supplementary Table 1). The SNP sets with the highest degree of similarity are S1_KO and S2_KO (jaccard index = 0.53). The Het sample which has altered global expression that resembles Opn4 KO expression (S5_Het) does have more SNPs shared with the Opn4 KO samples than its Opn4 Het counterparts. A majority of the SNPs called were sample-specific and not shared between like-genotyped samples (Supplementary Figure 5.3), though the S5_Het sample SNP profile was closer in relation to the KO samples than its Het counterparts based on the Euclidean distance (Supplementary Figure 5.4).

Further interrogation of the genetic drift in this mouse line is needed to determine regulatory factors influencing Opn4 expression. Genotyping Opn4 Het mice beyond coding regions would help identify potential genetic drift within this mouse line and possibly rule out single nucleotide polymorphisms (SNPs) as the root cause of Opn4 expression variability. Our data thus far suggests that researchers use caution when designing studies with this line. In addition to the Opn4 expression heterogeneity, recent studies have also outlined the non-specificity of the $Opn4^{CRE}$ mouse line when considering labeled cell types. Maloney et al. recently found variability in reporter expression depending on the specific reporter mouse line cross (Maloney et al., 2022). For example, two reporter mouse lines, the Ai9 and Ai14, elicit off-target labeling of non-ipRGC cells. Other Cre-reporter lines were found to under-label the ipRGC populations. Further investigation is needed to understand the Opn4 expression variability and the targeted population of cells using the $Opn4^{CRE}$ crossed with the $Rpl22^{HA}$ mouse lines.

5.6 Materials and Methods

All experimental data for this study was collected by Rashmi Gupta and Shyrice M. Mitchell. All data preprocessing and downstream analysis was performed by Theresa A Alexander. Figures were produced by Theresa A Alexander and edited by Theresa A Alexander and Rashmi Gupta.

5.6.1 Mouse lines

The Opn4-Cre transgenic mouse line (JAX stock #035925, Ecker et al., 2010), Ribotag mice (#029977, B6J.129(Cg) Rpl22<tm1.1Psam>/SjJ, Sanz E, et al. 2009) and Ai9(RCL-tdT) (B6.Cg-*Gt(ROSA)26Sortm9(CAG-tdTomato)Hze*/J) (JAX stock #007909, Madisen. et al., 2010) were obtained from Jackson. The transgenic line ($Rpl22^{HA}$) was crossed with the melanopsin mutant $Opn4^{CRE}$ transgenic line to produce a colony of mice that generates HA epitope-tagged ribosomes only in the presence of Cre recombinase. The three different genotypes generated for the study were $Opn4^{Cre/Cre}$::$Rpl22^{HA/+}$ (melanopsin knockout (KO) animal), $Opn4^{Cre/+}$::$Rpl22^{HA/+}$ (melanopsin expressing animals), and $Opn4^{+/+}$::$Rpl22^{HA/+}$(wild type animals). There was no phenotypic difference between $Opn4^{Cre/+}$ and $Opn4^{+/+}$ wild-type animals.

The tail tips of animals were collected on the same day as tissue collection. DNA was extracted using DNAeasy Blood and tissue kit (Qiagen, 69504), and PCR was performed using primers from published protocols (Primers for melanopsin are Forward: GGGTTCTGAGAGTGAAGTGG, Reverse: AAGAGGCCTTGAGTTCTCC (Ecker et al., 2010). Primers for CRE are

GCCAGCTAAACATGCTTCATC and ATTGCCCCTGTTTCACTATCC (Zhang et al 2007) Primers for Rpl22HA Forward: GGGAGGCTTGCT GGATATG Reverse: TTCCAGACACAGGCTAAGTACAC (JAX Laboratory). Some tail tips were sent to Transnetyx for genotyping. The Cre-negative samples will act as a negative control for non-specific binding during the TRAP experiment. All samples collected will be $Rpl22^{HA/+}$ as it was written in the published protocol (Shigeoka et al., 2018). In the current study, only male mice were used in order to minimize the variability in our data due to sex. At maturity, mice show sex-dependent differences in circadian regulation (Krizo & Mintz, 2015). During the study animals were maintained under standard conditions on a 12-hour day / 12- hour night cycle to standardize light exposure between experimental and control groups and all procedures were performed during the mouse light cycle between 1:00 to 5:00 PM to control for circadian rhythm. All animal procedures were approved and conducted under the guidelines of DLAR, University of Maryland, College Park.

5.6.2 Tissue Collection

In the current study as I was interested in melanopsin-dependent local protein translation and ipRGC activity is developmentally regulated hence during the study, the tissue samples were collected at two different developmental time points, I) before eye-opening at P8 when the activity is photoreceptor independent i.e. governed by melanopsin and, II) after eye-opening at P15 when the activity is photoreceptor dependent. Exactly 12 hours before tissue collection, animals were anesthetized with isofluorane and 2ul of 200ug/ml CTß-Alexa488 was bilaterally injected into the eyes

of the animal. The next day animals were anesthetized using isofluorane and quickly decapitated. Eyes and brain were quickly submerged into ~ 5 ml of fresh 200 µM cycloheximide PBS ice slush to prevent ribosome runoff. Brains were sliced into 750uM sections on the vibratome in the presence of 200 µM cycloheximide PBS ice slush. The targets (SCN and dLGN) from brain slices and retinae from eyes were manually dissected and collected into separate low-bind Eppendorf tubes, and quickly frozen in liquid nitrogen within 30 minutes to prevent ribosome runoff and stored at −80°C until tissue homogenization.

5.6.3 Axon- Translation ribosome affinity purification (TRAP)

In order to directly retrieve axonal translatome the immunoprecipitation of ribosome-associated mRNA was performed according to published protocols (Shigeoka et al., 2016, 2018). After genotyping the tissues from the same age, genotype and tissue type were pooled together (7 Retinas, dLGN, and SCN for one group) for TRAP (Translating ribosome affinity purification). Tissues were homogenized using a hand-held pellet pestle tissue grinder 3-4 times at an interval of 15-20 sec in lysis buffer (20mM HEPES-KOH, 5mM MgCl2, 150mM KCl, 1mM DTT, SUPERase In, and Complete EDTA-free Protease Inhibitor Cocktail) in the presence of cycloheximide (200uM) (to stop translational elongation and to lock translating ribosomes on the mRNA) on ice. The homogenates were centrifuged at 16,000g at 4 °C for 10 mins. The supernatant collected was pre-cleared using 80 ul of Protein G magnetic beads (Thermo Scientific™ Pierce™ #88848). The tubes

were rotated head to toe at 4 °C for 1 hour. The tube containing beads with lysate was placed on a magnetic stand and the supernatant was collected in a new Eppendorf tube. TRAP protocol was optimized before performing axon-TRAP using the polyclonal anti-HA antibody (Abcam, ab9110). During the TRAP protocol, precleared ribosome-mRNA complexes were immunoprecipitated for overnight incubation using 5ul of polyclonal anti-HA antibody followed by incubation for four hours using 80 ul of Dynabeads Protein G (Life Technologies, 10004D). After four washes with lysis buffer, the beads were suspended in 100ul of CHX Lysis buffer.

5.6.4 Axon-TRAP-RiboTag RNA Isolation

The conjugated beads in the lysis buffer were incubated with Trizol (500ul) (Invitrogen, 15596026) at room temperature for 5 mins. The conjugated beads RNA was vortexed and centrifuged for 1min and placed on a magnetic stand for 30 sec. The supernatant containing trizol reagent was transferred to another tube and mixed with 100 ul (5:1) of chloroform. The lysate mix was mixed by inverting the tube and kept for 2-3 min of incubation on a stand, followed by centrifugation (12,000 g) at 4 °C for 15 mins. The aqueous phase collected was mixed with 250ul of isopropanol and centrifuged (10,000 g) at 4 °C for 10 mins. The supernatant obtained was aspirated and washed using 500ul of 75% ethanol followed by DNase I treatment (Qiagen) to remove genomic DNA contamination. Finally, the samples were eluted in 14ul of RNase-free water. The RNA samples were checked for quantity using Qubit 3 flourometer High Sensitivity RNA Assay Kit (Invitrogen™, Q32852) and

RNA integrity was determined using the Agilent Bioanalyzer 4150 (Agilent Technologies) using the HS RNA assay kit (5067). The RNA isolated was stored at -80 °C before sending it to the BBI core.

5.6.5 Axon-TRAP-RiboTag Library preparation and sequencing

RNA library preparation from TRAP samples of melanopsin WT, Het, and KO was performed at *Brain & Behavior Institute - Advanced Genomic Technologies Core (BBI-AGTC*). Briefly, RNASeq libraries were made using TakaraBio SMARTer Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian Kit (Takara, 634486) with SMARTer RNA Unique Dual Index Kit – 96U (8bp) (384 reactions) (Takara, 634452) for Illumina Sequencing. Briefly, the library was prepared using 1–6 ng of total RNA with a fragmentation time of 2 mins, 5 cycles of amplification in PCR1, and depending on the concentration of RNA the amplification cycles for PCR2 vary between 12-14 cycles. Further, the quality of the cDNA library prepared was analyzed on Agilent Tape station 4150 using the HSD1000 Assay kit. Libraries were pooled and run on an Illumina sequencer (Illumina NextSeq 1000 (SNVL00136) having a 60 bp paired-end read protocol. A single pool was prepared by combining 10ul of each sample at a 2 nM dilution. Illumina uses BCL convert to generate FASTQ files from binary base call (BCL) sequence files. The raw sequencing data were received in FASTQ format.

5.6.6 Axon-TRAP-RiboTag RNA-sequencing read preprocessing

Fastq files from axon-TRAPed samples were de-multiplexed using the Cogent NGS analysis pipeline (CogentAP) from Takara Bio (v1.0). The "cogent demux" function was used to assign the reads to the samples based on the sample barcodes provided in the custom well-list files. The "cogent analyze" function was used to perform read trimming with cutadapt (version 3.2) (M. Martin, 2011) to trim sequencing adapters, unique molecular identifiers (UMIs), and any low-quality read sequences defined by a low base pair call score. After trimming each sequenced read, the result will be high quality biological sequence. Trimmed reads were then mapped to the Genome Reference Consortium Mouse Build 38 (Schneider et al., 2016) using HISAT2 (D. Kim et al., 2019) and gene counts were obtained using HTSeq (Anders et al., 2015; Putri et al., 2022).

Sample inclusion was determined by visually inspecting the PCA projections, as well as the number of nonzero genes observed criteria described in Chapter 4. Final sample numbers reported by each experimental condition in Table 5.4. Conditions which were included in downstream analysis included the retina (all genotypes) and the dLGN (all genotypes), while SCN samples were only used for PCA and nonzero numbers of genes analyses. Genes were filtered out of the dataset with <= 10 counts across all samples (13,923 remaining genes after filtering). Surrogate variable analysis (SVA) (Leek & Storey, 2007) was used for batch correction to correct for

variational differences between experimental batches (experimental batch 1 n = 2, experimental batch 2 n = 15).

Table 5.4. Sample numbers for P15 axon-TRAP experiment.

axon-TRAP RNA-seq Sample Numbers			
	Retina	SCN	dLGN
Opn4 Het	2	1	2
Opn4 KO	2	1	2
Opn4 WT (genotype for background correction)	3	1	3

5.6.7 Tissue Collection for Bulk RNA sequencing (RNA-seq)

The tissue procurement protocol for bulk RNA sequencing was the same as for TRAP experiments. Briefly, after dissection, the eyes and brain were quickly submerged into ~ 5 ml of fresh 1X PBS. Brains were sliced into 750uM sections on the vibratome in the presence of 1X PBS ice slush. The targets SCN and dLGN and retinae from eyes were manually dissected and collected into separate low-bind Eppendorf tubes, and quickly frozen in liquid nitrogen within 30 minutes and stored at −80°C until tissue homogenization.

5.6.8 RNA isolation for Bulk RNA-seq

Tissues were homogenized using a pellet pestle hand-held tissue grinder in lysis buffer (20mM HEPES-KOH, 5mM MgCl2, 150mM KCl, 1mM DTT, SUPERase In, and Complete EDTA-free Protease Inhibitor Cocktail). The homogenates were

centrifuged at 16,000g at 4 °C for 10 mins. The supernatant obtained was incubated while mixing the tube by inverting with 500ul of ice-cold Trizol (Invitrogen, 15596026) at room temperature for 5 mins, and 100 ul (5:1) of chloroform was added to the lysate. After 2-3 min of incubation, the tube was vortexed and centrifuged for 15 mins at 12,000 g at 4 °C. The aqueous phase obtained was then transferred to another tube, mixed with 250ul of isopropanol, and centrifuged (10,000 g) at 4 °C for 10 mins. After centrifugation, the supernatant obtained was aspirated to another tube and 500ul of 75% ethanol was added. The tubes were centrifuged at 7500g for 5 min. RNA purification was performed according to manufacturer instructions using DNase I (Qiagen). Finally, the samples were eluted in 15ul of RNase-free water. RNA quantity was measured using Qubit 3 fluorometer High Sensitivity RNA Assay Kit (Invitrogen™, Q32852), and RNA integrity was determined using the Agilent Tape station 4150 (Agilent Technologies) using the RNA and HS RNA assay kit. The RNA isolated was stored at -80 °C before sending it to the BBI core.

5.6.9 Library preparation and sequencing of Bulk RNA-seq

Total RNA isolated from melanopsin Het, and KO was used for RNAseq experiments. Library preparation and RNAseq were performed at *Brain & Behavior Institute - Advanced Genomic Technologies Core (BBI-AGTC*). Briefly, RNA-seq libraries were made using Illumina Stranded mRNA with IDT for Illumina DNA/RNA UD Indexes Sets A for Illumina Sequencing. The library for RNAseq was prepared based on adenylation of 3’ ends by using 275 ng of total RNA to a final volume of 25ul using

RNase-free water. The protocol used includes a fragmentation time of 8 mins and 12 cycles of amplification in PCR. Further, the quality of the cDNA library prepared was analyzed on Agilent Tape station 4150 using the D1000 Assay kit. Following library preparation using PolyA selection, the concentration of samples was determined using qPCR from Roche Lightcycler 96. Libraries were pooled and processed in a single batch run on an Illumina sequencer (Illumina NextSeq 1000 (SNVL00136) having a 58 bp paired-end read protocol (dual index). A single pool was prepared by combining 10ul of each sample at a 2 nM dilution. The raw sequencing data were received in FASTQ format.

5.6.10 Bulk RNA-sequencing read preprocessing

Fastq files from bulk-RNA sequenced samples were trimmed using trimmomatic (Bolger et al., 2014) to remove illumina and epicenter adapters and perform a sliding window removal of any base with a quality score basecall < 25. Trimmed reads were then mapped to the Genome Reference Consortium Mouse Build 38 (Schneider et al., 2016) using HISAT2 (D. Kim et al., 2019) and gene counts were obtained using HTSeq (Anders et al., 2015; Putri et al., 2022).

Opn4 expression was found to be variable even within the like-genotype Opn4 hets (Supplementary Figure 5.1). Sample filtering was conducted to include only samples with CPM normalized Opn4 expression > 10 CPMs for Opn4 hets and < .25 CPMs per genotype (n = 15 samples post-filtering) (Table 5.5), and genes were filtered out

of the dataset with <= 10 counts across all samples (14,107 remaining genes after filtering).

Table 5.5. Sample numbers for P8 bulk RNA-seq experiment.

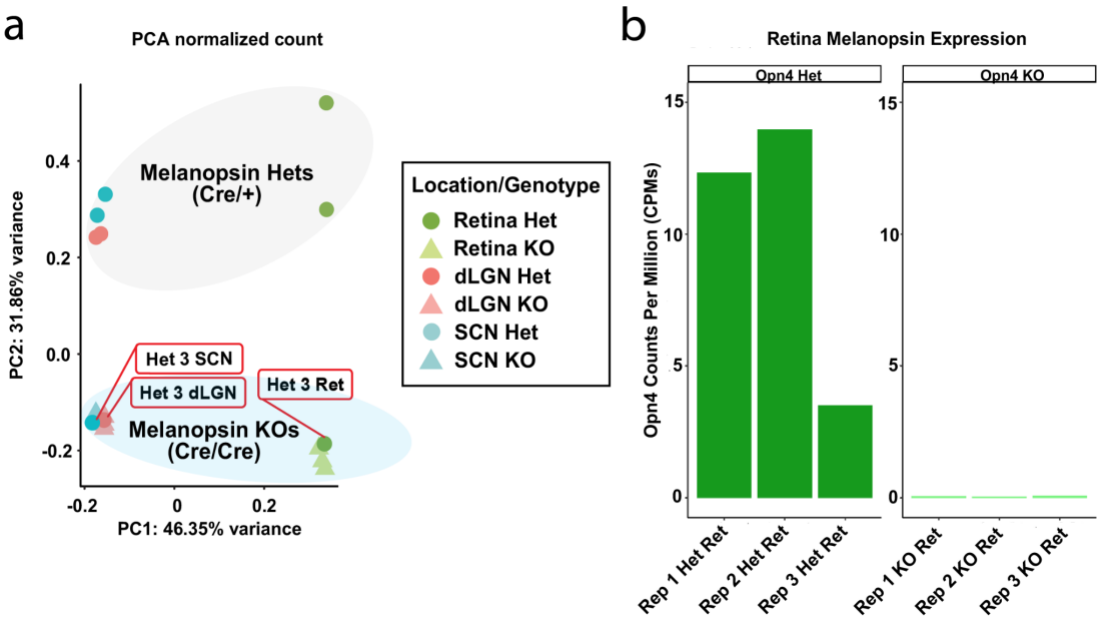
Bulk RNA-seq Pre-filtering Sample Numbers			
	Retina	SCN	dLGN
Opn4 Het	3	3	3
Opn4 KO	3	3	3
Bulk RNA-seq Post-filtering Sample Numbers			
	Retina	SCN	dLGN
Opn4 Het	2	2	2
Opn4 KO	3	3	3

5.6.11 Bioinformatic Analyses

Downstream bioinformatic analyses were completed using R version 4.2.1 (R Core Team, 2021) using custom functions developed by Theresa Alexander as well as functions developed by Ashton Trey Belew as part of the hpgltools package (https://github.com/elsayed-lab/hpgltools). Lowly-expressed genes were filtered to include only those with > 10 reads across all samples (14,107 remaining genes after filtering). Count data was converted to counts per million (CPMs), then quantile normalized, and log2 transformed before principal component analysis (PCA). Differential expression (DE) analysis was conducted using the raw count data as input to the DESeq2 R package (Love et al., 2014). All contrasts completed for the

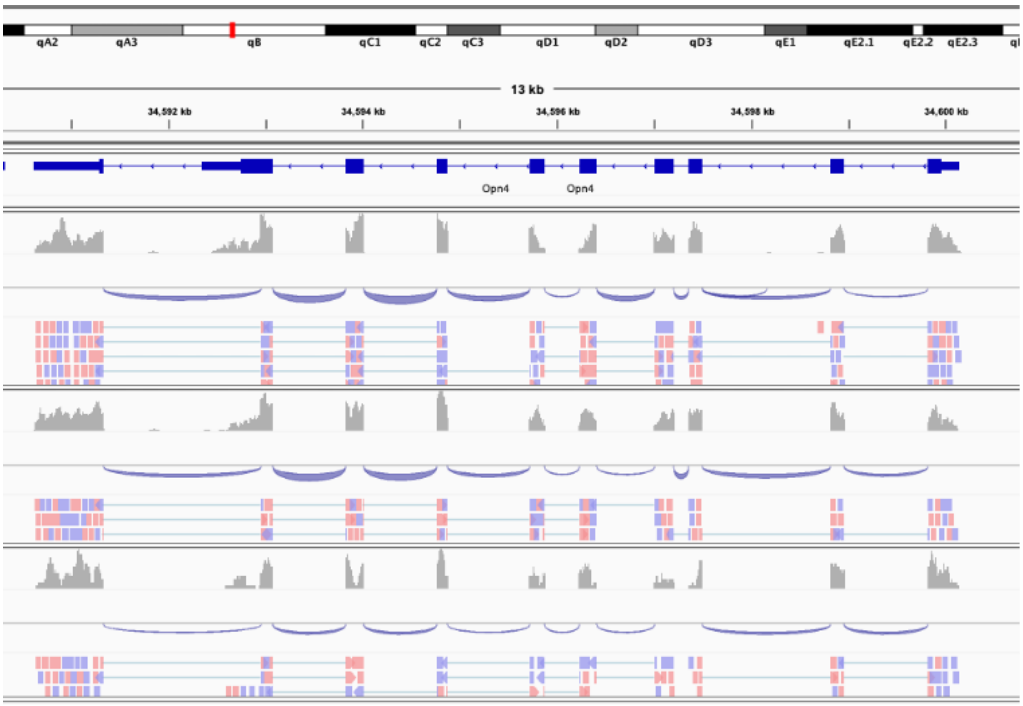
bulk RNA-sequencing data have full triplicate samples for each experimental condition. Contrasts completed for the axon-TRAP-RiboTag data are not sufficiently powered, with n = 2 for each experimental condition (see Chapter 4 for explanation of sample inclusion criteria leading to such few samples in these contrasts). At the time of this writing, additional samples for this experiment are being collected such that these contrasts will be fully powered. The DE analyses for the translatomic results are all preliminary and are subject to change with additional samples added to the study. Significance thresholds for all DE analysis were set as the magnitude of log2 fold change > 1 and adjusted *P* value <= 0.05. The R package gProfiler2 (Kolberg et al., 2020) was used to conduct gene set enrichment analysis (GSEA) with gene set annotation databases Gene Ontology (GO), KEGG, and REACTOME. An adjusted *P* value cutoff of <= 0.05 was used for GSEA.

5.7 Supplementary Material



Supplementary Figure 5.1. Heterogeneity in Opn4 expression in bulk RNA-seq.

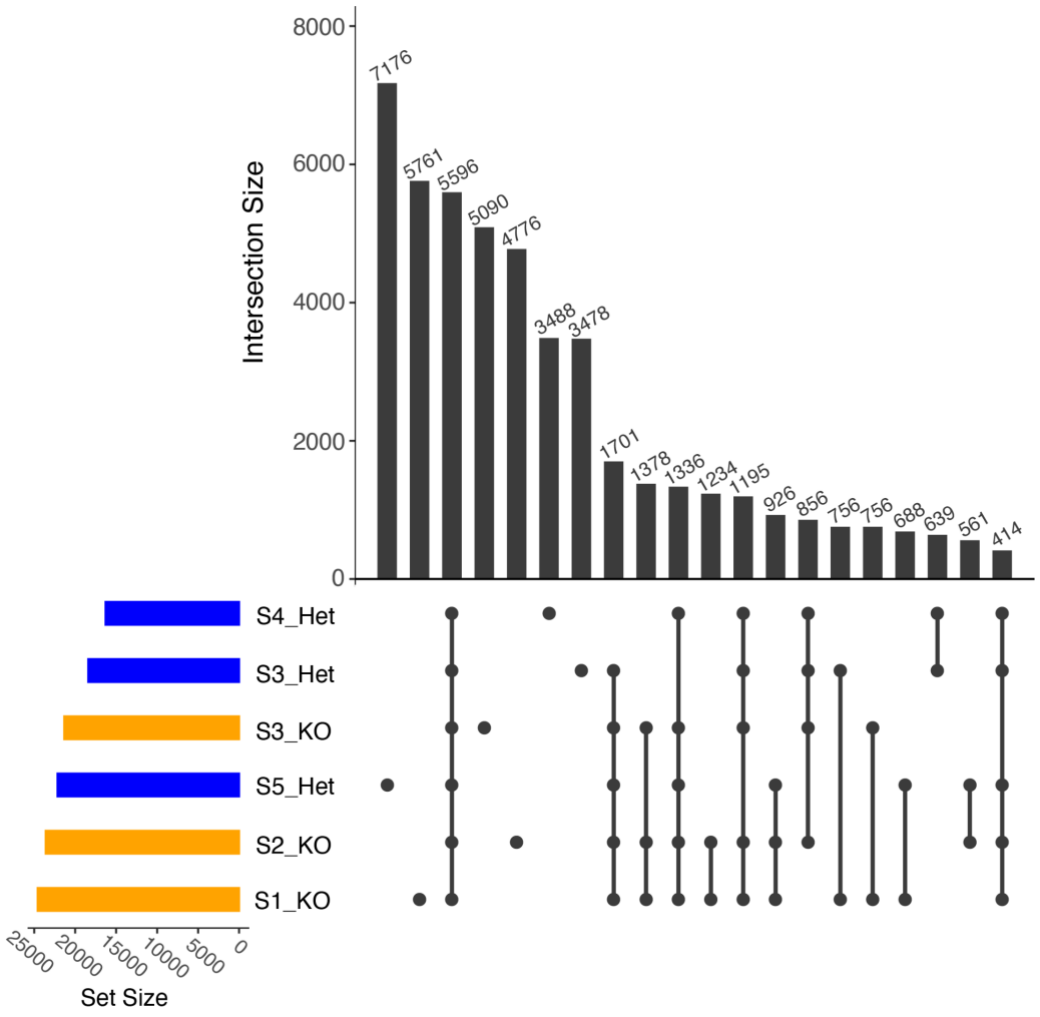
Melanopsin expression is not consistent within Opn4 Het replicates. a) Principal component analysis (PCA) of all bulk RNA-sequenced samples shows discrete clustering with respect to location and genotype with the exception of 3 samples (all collected from the same mouse). b) Comparison of the CPM normalized Opn4 expression in the retina Opn4 het samples shows decreased expression of Opn4 in replicate number 3, despite having the same Opn4 genotype.



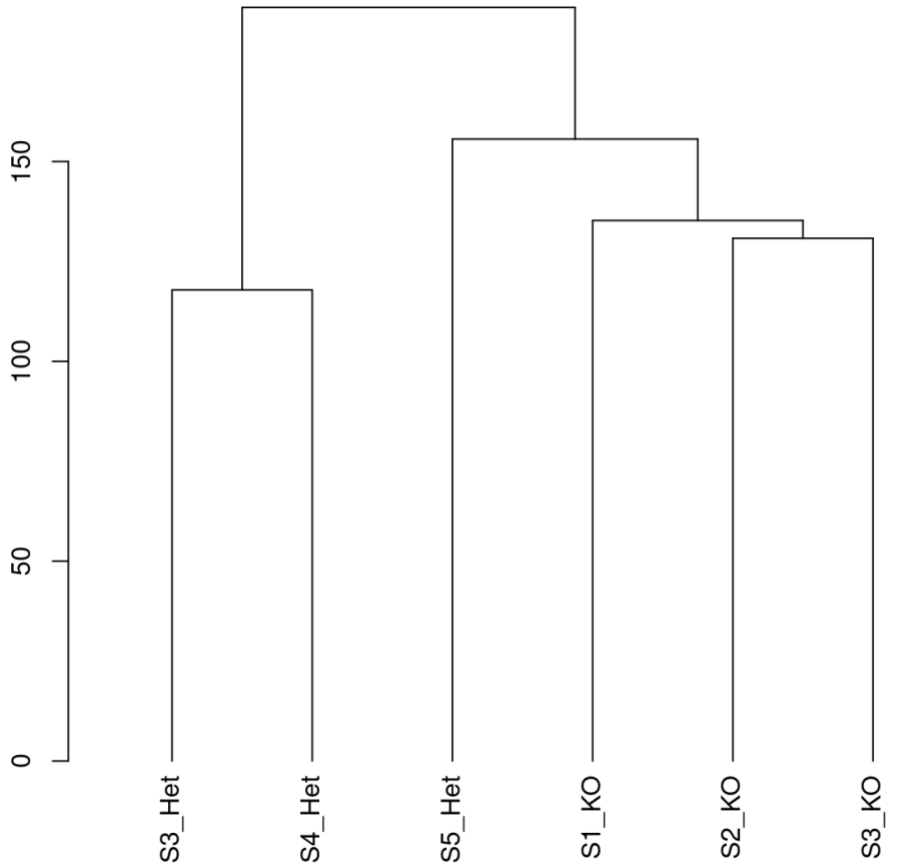
Supplementary Figure 5.2. IGV view of Opn4 aligned reads.
Integrative Genome Viewer (IGV) snapshot of the aligned transcriptome sequencing reads to the mus musculus v38 genome Opn4 coding sequence (CDS) region for three Opn4 Het mouse samples. There is no evident difference in Opn4 isoform expression between the samples as well as not polymorphic nucleotides in this region.

Supplementary Table 1. Jaccard index of sample SNPs.
Variants were called for each bulk RNA-seq sample out of the 58,075 SNPs called from the transcriptome.

	S1_KO	S2_KO	S3_KO	S3_Het	S4_Het	S5_Het
S1_KO	1					
S2_KO	0.53	1				
S3_KO	0.48	0.51	1			
S3_Het	0.30	0.35	0.34	1		
S4_Het	0.27	0.33	0.32	0.42	1	
S5_Het	0.42	0.44	0.40	0.28	0.26	1



Supplementary Figure 5.3. Comparison of SNP sets between bulk RNA-seq samples. Upset plot of shared SNP sets across all retina bulk RNA-seq samples. The overall set size for each group for the number of SNPs within each sample is shown in the histogram to the left (blue = Opn4 Het sample, orange = Opn4 KO sample). The top histogram indicates the number of SNPs in each set (set group inclusion shown with line-dots below histogram).



Supplementary Figure 5.4. Dendrogram of clustered bulk RNA-seq samples.
The euclidean distance between SNP profiles of each bulk RNA-seq sample were hierarchically clustered and plotted as a dendrogram.

Bibliography

Adams, R. H., Wilkinson, G. A., Weiss, C., Diella, F., Gale, N. W., Deutsch, U., Risau, W., & Klein, R. (1999). Roles of ephrinB ligands and EphB receptors in cardiovascular development: Demarcation of arterial/venous domains, vascular morphogenesis, and sprouting angiogenesis. *Genes & Development*, *13*(3), 295–306. https://doi.org/10.1101/gad.13.3.295

Afzal, A., Shaw, L. C., Ljubimov, A. V., Boulton, M. E., Segal, M. S., & Grant, M. B. (2007). Retinal and choroidal microangiopathies: Therapeutic opportunities. *Microvascular Research*, *74*(2–3), 131–144. https://doi.org/10.1016/j.mvr.2007.04.011

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Andrae, J., Gallini, R., & Betsholtz, C. (2008). Role of platelet-derived growth factors in physiology and medicine. *Genes & Development*, *22*(10), 1276–1312. https://doi.org/10.1101/gad.1653708

Andreassi, C., Crerar, H., & Riccio, A. (2018). Post-transcriptional Processing of mRNA in Neurons: The Vestiges of the RNA World Drive Transcriptome Diversity. *Frontiers in Molecular Neuroscience*, *11*, 304. https://doi.org/10.3389/fnmol.2018.00304

Arroyo, D. A., Kirkby, L. A., & Feller, M. B. (2016). Retinal Waves Modulate an Intraretinal Circuit of Intrinsically Photosensitive Retinal Ganglion Cells. *Journal of Neuroscience*, *36*(26), 6892–6905.

https://doi.org/10.1523/JNEUROSCI.0572-16.2016

Auer, P. L., & Doerge, R. W. (2010). Statistical Design and Analysis of RNA Sequencing Data. *Genetics*, *185*(2), 405–416. https://doi.org/10.1534/genetics.110.114983

Bartley, C. M., O'Keefe, R. A., Blice-Baum, A., Mihailescu, M.-R., Gong, X., Miyares, L., Karaca, E., & Bordey, A. (2016). Mammalian FMRP S499 Is Phosphorylated by CK2 and Promotes Secondary Phosphorylation of FMRP. *Eneuro*, *3*(6), ENEURO.0092-16.2016. https://doi.org/10.1523/ENEURO.0092-16.2016

Bass, A. J., Robinson, D. G., & Storey, J. D. (2019). *Determining sufficient sequencing depth in RNA-Seq differential expression studies* [Preprint]. Genomics. https://doi.org/10.1101/635623

Beier, C., Zhang, Z., Yurgel, M., & Hattar, S. (2020). Projections of ipRGCs and conventional RGCs to retinorecipient brain nuclei. *Journal of Comparative Neurology*, cne.25061. https://doi.org/10.1002/cne.25061

Berg, D. J., Kartheiser, K., Leyrer, M., Saali, A., & Berson, D. M. (2019). Transcriptomic Signatures of Postnatal and Adult Intrinsically Photosensitive Ganglion Cells. *Eneuro*, *6*(4), ENEURO.0022-19.2019. https://doi.org/10.1523/ENEURO.0022-19.2019

Berson, D. (2003). Strange vision: Ganglion cells as circadian photoreceptors. *Trends in Neurosciences*, *26*(6), 314–320. https://doi.org/10.1016/S0166-2236(03)00130-9

Berson, D. M. (2002). Phototransduction by Retinal Ganglion Cells That Set the

Circadian Clock. *Science*, *295*(5557), 1070–1073. https://doi.org/10.1126/science.1067262

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brady, S., Colman, D. R., & Brophy, P. (2014). Subcellular Organization of the Nervous System. In *From Molecules to Networks* (pp. 23–52). Elsevier. https://doi.org/10.1016/B978-0-12-397179-1.00002-6

Brown, B. C., Bray, N. L., & Pachter, L. (2018). Expression reflects population structure. *PLOS Genetics*, *14*(12), e1007841. https://doi.org/10.1371/journal.pgen.1007841

Bush, S. J., Freem, L., MacCallum, A. J., O'Dell, J., Wu, C., Afrasiabi, C., Psifidi, A., Stevens, M. P., Smith, J., Summers, K. M., & Hume, D. A. (2018). Combination of novel and public RNA-seq datasets to generate an mRNA expression atlas for the domestic chicken. *BMC Genomics*, *19*(1), 594. https://doi.org/10.1186/s12864-018-4972-7

Caval-Holme, F. S., Aranda, M. L., Chen, A. Q., Tiriac, A., Zhang, Y., Smith, B., Birnbaumer, L., Schmidt, T. M., & Feller, M. B. (2022). The Retinal Basis of Light Aversion in Neonatal Mice. *The Journal of Neuroscience*, *42*(20),

4101–4115. https://doi.org/10.1523/JNEUROSCI.0151-22.2022

Choe, H. K., & Cho, J. (2020). Comprehensive Genome-Wide Approaches to Activity-Dependent Translational Control in Neurons. *International Journal of Molecular Sciences*, *21*(5), 1592. https://doi.org/10.3390/ijms21051592

Clara, C. A., Marie, S. K. N., de Almeida, J. R. W., Wakamatsu, A., Oba-Shinjo, S. M., Uno, M., Neville, M., & Rosemberg, S. (2014). Angiogenesis and expression of PDGF-C, VEGF, CD105 and HIF-1α in human glioblastoma: Angiogenesis and glioblastoma. *Neuropathology*, n/a-n/a. https://doi.org/10.1111/neup.12111

Dalla Costa, I., Buchanan, C. N., Zdradzinski, M. D., Sahoo, P. K., Smith, T. P., Thames, E., Kar, A. N., & Twiss, J. L. (2021). The functional organization of axonal mRNA transport and translation. *Nature Reviews Neuroscience*, *22*(2), 77–91. https://doi.org/10.1038/s41583-020-00407-7

Davies, M. H., Stempel, A. J., Hubert, K. E., & Powers, M. R. (2010). Altered vascular expression of EphrinB2 and EphB4 in a model of oxygen-induced retinopathy. *Developmental Dynamics*, *239*(6), 1695–1707. https://doi.org/10.1002/dvdy.22306

Dell, S., Peters, S., Mu¨ther, P., Kociok, N., & Joussen, A. M. (2006). The Role of PDGF Receptor Inhibitors and PI3-Kinase Signaling in the Pathogenesis of Corneal Neovascularization. *Investigative Opthalmology & Visual Science*, *47*(5), 1928. https://doi.org/10.1167/iovs.05-1071

Ding, C., & He, X. (2004). *K* -means clustering via principal component analysis. *Twenty-First International Conference on Machine Learning - ICML '04*, 29.

https://doi.org/10.1145/1015330.1015408

Do, M. T. H. (2019). Melanopsin and the Intrinsically Photosensitive Retinal Ganglion Cells: Biophysics to Behavior. *Neuron*, *104*(2), 205–226. https://doi.org/10.1016/j.neuron.2019.07.016

Do, M. T. H., & Yau, K.-W. (2010). Intrinsically Photosensitive Retinal Ganglion Cells. *Physiological Reviews*, *90*(4), 1547–1581. https://doi.org/10.1152/physrev.00013.2010

Donovan, J., Shiwen, X., Norman, J., & Abraham, D. (2013). Platelet-derived growth factor alpha and beta receptors have overlapping functional activities towards fibroblasts. *Fibrogenesis & Tissue Repair*, *6*(1), 10. https://doi.org/10.1186/1755-1536-6-10

Dougherty, J. D. (2017). The Expanding Toolkit of Translating Ribosome Affinity Purification. *The Journal of Neuroscience*, *37*(50), 12079–12087. https://doi.org/10.1523/JNEUROSCI.1929-17.2017

Ecker, J. L., Dumitrescu, O. N., Wong, K. Y., Alam, N. M., Chen, S.-K., LeGates, T., Renna, J. M., Prusky, G. T., Berson, D. M., & Hattar, S. (2010). Melanopsin-Expressing Retinal Ganglion-Cell Photoreceptors: Cellular Diversity and Role in Pattern Vision. *Neuron*, *67*(1), 49–60. https://doi.org/10.1016/j.neuron.2010.05.023

Fernandopulle, M. S., Lippincott-Schwartz, J., & Ward, M. E. (2021). RNA transport and local translation in neurodevelopmental and neurodegenerative disease. *Nature Neuroscience*, *24*(5), 622–632. https://doi.org/10.1038/s41593-020-00785-2

Fu, Y., Zhong, H., Wang, M.-H. H., Luo, D.-G., Liao, H.-W., Maeda, H., Hattar, S., Frishman, L. J., & Yau, K.-W. (2005). Intrinsically photosensitive retinal ganglion cells detect light with a vitamin A-based photopigment, melanopsin. *Proceedings of the National Academy of Sciences*, *102*(29), 10339–10344. https://doi.org/10.1073/pnas.0501866102

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing*. https://doi.org/10.48550/ARXIV.1207.3907

Glock, C., Biever, A., Tushev, G., Nassim-Assir, B., Kao, A., Bartnik, I., tom Dieck, S., & Schuman, E. M. (2021). The translatome of neuronal cell bodies, dendrites, and axons. *Proceedings of the National Academy of Sciences*, *118*(43), e2113929118. https://doi.org/10.1073/pnas.2113929118

Guedes-Dias, P., & Holzbaur, E. L. F. (2019). Axonal transport: Driving synaptic function. *Science*, *366*(6462), eaaw9997. https://doi.org/10.1126/science.aaw9997

Hafner, A.-S., Donlin-Asp, P. G., Leitch, B., Herzog, E., & Schuman, E. M. (2019). Local protein synthesis is a ubiquitous feature of neuronal pre- and postsynaptic compartments. *Science*, *364*(6441), eaau3644. https://doi.org/10.1126/science.aau3644

Harris, J. J., Jolivet, R., & Attwell, D. (2012). Synaptic Energy Use and Supply. *Neuron*, *75*(5), 762–777. https://doi.org/10.1016/j.neuron.2012.08.019

Hattar, S. (2002). Melanopsin-Containing Retinal Ganglion Cells: Architecture, Projections, and Intrinsic Photosensitivity. *Science*, *295*(5557), 1065–1070. https://doi.org/10.1126/science.1069609

Heiman, M., Schaefer, A., Gong, S., Peterson, J. D., Day, M., Ramsey, K. E., Suárez-Fariñas, M., Schwarz, C., Stephan, D. A., Surmeier, D. J., Greengard, P., & Heintz, N. (2008). A Translational Profiling Approach for the Molecular Characterization of CNS Cell Types. *Cell*, *135*(4), 738–748. https://doi.org/10.1016/j.cell.2008.10.028

Hoffman, G. E., & Schadt, E. E. (2016). variancePartition: Interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics*, *17*(1), 483. https://doi.org/10.1186/s12859-016-1323-z

Holt, C. E., & Schuman, E. M. (2013). The Central Dogma Decentralized: New Perspectives on RNA Function and Local Translation in Neurons. *Neuron*, *80*(3), 648–657. https://doi.org/10.1016/j.neuron.2013.10.036

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jung, H., Yoon, B. C., & Holt, C. E. (2012). Axonal mRNA localization and local protein synthesis in nervous system assembly, maintenance and repair. *Nature Reviews Neuroscience*, *13*(5), 308–324. https://doi.org/10.1038/nrn3210

Kaczmarek, R., Gajdzis, P., & Gajdzis, M. (2021). Eph Receptors and Ephrins in Retinal Diseases. *International Journal of Molecular Sciences*, *22*(12), 6207. https://doi.org/10.3390/ijms22126207

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Kim, K.-Y., Rios, L. C., Le, H., Perez, A. J., Phan, S., Bushong, E. A., Deerinck, T. J., Liu, Y. H., Ellisman, M. A., Lev-Ram, V., Ju, S., Panda, S. A., Yoon, S., Hirayama, M., Mure, L. S., Hatori, M., Ellisman, M. H., & Panda, S. (2019). Synaptic Specializations of Melanopsin-Retinal Ganglion Cells in Multiple Brain Regions Revealed by Genetic Label for Light and Electron Microscopy. *Cell Reports*, *29*(3), 628-644.e6. https://doi.org/10.1016/j.celrep.2019.09.006

Kim, L. A., & D'Amore, P. A. (2012). A Brief History of Anti-VEGF for the Treatment of Ocular Angiogenesis. *The American Journal of Pathology*, *181*(2), 376–379. https://doi.org/10.1016/j.ajpath.2012.06.006

Kim, T., Chen, I. R., Lin, Y., Wang, A. Y.-Y., Yang, J. Y. H., & Yang, P. (2019). Impact of similarity metrics on single-cell RNA-seq data clustering. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bby076

Kirkby, L. A., & Feller, M. B. (2013). Intrinsically photosensitive ganglion cells contribute to plasticity in retinal wave circuits. *Proceedings of the National Academy of Sciences*, *110*(29), 12090–12095. https://doi.org/10.1073/pnas.1222150110

Kiselev, V. Y., Andrews, T. S., & Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, *20*(5), 273–282. https://doi.org/10.1038/s41576-018-0088-9

Kleiveland, C. R. (2015). Peripheral Blood Mononuclear Cells. In K. Verhoeckx, P. Cotter, I. López-Expósito, C. Kleiveland, T. Lea, A. Mackie, T. Requena, D. Swiatecka, & H. Wichers (Eds.), *The Impact of Food Bioactives on Health* (pp. 161–167). Springer International Publishing. https://doi.org/10.1007/978-

3-319-16104-4_15

Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J., & Peterson, H. (2020). gprofiler2—An R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Research*, *9*, 709. https://doi.org/10.12688/f1000research.24956.2

Krizo, J. A., & Mintz, E. M. (2015). Sex Differences in Behavioral Circadian Rhythms in Laboratory Rodents. *Frontiers in Endocrinology*, *5*. https://doi.org/10.3389/fendo.2014.00234

Krock, B. L., Skuli, N., & Simon, M. C. (2011). Hypoxia-Induced Angiogenesis: Good and Evil. *Genes & Cancer*, *2*(12), 1117–1133. https://doi.org/10.1177/1947601911423654

Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., … Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, *501*(7468), 506–511. https://doi.org/10.1038/nature12531

Laurens van der Maaten & Geoffrey Hinton. (2008). Visulatizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Lazzerini Ospri, L., Prusky, G., & Hattar, S. (2017). Mood, the Circadian System, and Melanopsin Retinal Ganglion Cells. *Annual Review of Neuroscience*, *40*(1), 539–556. https://doi.org/10.1146/annurev-neuro-072116-031324

Lee, C. W., & Peng, H. B. (2008). The Function of Mitochondria in Presynaptic

Development at the Neuromuscular Junction. *Molecular Biology of the Cell*, *19*(1), 150–158. https://doi.org/10.1091/mbc.e07-05-0515

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews. Genetics*, *11*(10), 733–739. https://doi.org/10.1038/nrg2825

Leek, J. T., & Storey, J. D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genetics*, *3*(9), 12.

Lemmon, M. A., & Schlessinger, J. (2010). Cell Signaling by Receptor Tyrosine Kinases. *Cell*, *141*(7), 1117–1134. https://doi.org/10.1016/j.cell.2010.06.011

Leung, K.-M., van Horck, F. P., Lin, A. C., Allison, R., Standart, N., & Holt, C. E. (2006). Asymmetrical β-actin mRNA translation in growth cones mediates attractive turning to netrin-1. *Nature Neuroscience*, *9*(10), 1247–1256. https://doi.org/10.1038/nn1775

Lever, J., Krzywinski, M., & Altman, N. (2017). Points of Significance: Principal component analysis. *Nature Methods*, *14*, 641–642. https://doi.org/10.1038/nmeth.4346

Li, L., Yu, J., & Ji, S.-J. (2021). Axonal mRNA localization and translation: Local events with broad roles. *Cellular and Molecular Life Sciences*, *78*(23), 7379–7395. https://doi.org/10.1007/s00018-021-03995-4

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lucas, J. A., & Schmidt, T. M. (2019). Cellular properties of intrinsically photosensitive retinal ganglion cells during postnatal development. *Neural Development, 14*(1), 8. https://doi.org/10.1186/s13064-019-0132-2

Maday, S., Twelvetrees, A. E., Moughamian, A. J., & Holzbaur, E. L. F. (2014). Axonal Transport: Cargo-Specific Mechanisms of Motility and Regulation. *Neuron, 84*(2), 292–309. https://doi.org/10.1016/j.neuron.2014.10.019

Mamer, S. B., Chen, S., Weddell, J. C., Palasz, A., Wittenkeller, A., Kumar, M., & Imoukhuede, P. I. (2017). Discovery of High-Affinity PDGF-VEGFR Interactions: Redefining RTK Dynamics. *Scientific Reports, 7*(1), 16439. https://doi.org/10.1038/s41598-017-16610-z

Martin, K. C. (2004). Local protein synthesis during axon guidance and synaptic plasticity. *Current Opinion in Neurobiology, 14*(3), 305–310. https://doi.org/10.1016/j.conb.2004.05.009

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal, 17*(1), 10. https://doi.org/10.14806/ej.17.1.200

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv:1802.03426 [Cs, Stat]*. http://arxiv.org/abs/1802.03426

McNeill, D. S., Sheely, C. J., Ecker, J. L., Badea, T. C., Morhardt, D., Guido, W., & Hattar, S. (2011). Development of melanopsin-based irradiance detecting circuitry. *Neural Development, 6*(1), 8. https://doi.org/10.1186/1749-8104-6-8

Mersha, T. B., & Abebe, T. (2015). Self-reported race/ethnicity in the age of genomic

research: Its potential impact on understanding health disparities. *Human Genomics*, *9*(1). https://doi.org/10.1186/s40246-014-0023-x

Narayanan, U., Nalavadi, V., Nakamoto, M., Pallas, D. C., Ceman, S., Bassell, G. J., & Warren, S. T. (2007). FMRP Phosphorylation Reveals an Immediate-Early Signaling Pathway Triggered by Group I mGluR and Mediated by PP2A. *The Journal of Neuroscience*, *27*(52), 14349–14357. https://doi.org/10.1523/JNEUROSCI.2969-07.2007

O'Neal, W. T., Griffin, W. F., Dries-Devlin, J. L., Kent, S. D., Chen, J., Willis, M. S., & Virag, J. A. I. (2013). Ephrin–Eph signaling as a potential therapeutic target for the treatment of myocardial infarction. *Medical Hypotheses*, *80*(6), 738–744. https://doi.org/10.1016/j.mehy.2013.02.024

Ostroff, L. E., Santini, E., Sears, R., Deane, Z., Kanadia, R. N., LeDoux, J. E., Lhakhang, T., Tsirigos, A., Heguy, A., & Klann, E. (2019). Axon TRAP reveals learning-associated alterations in cortical axonal mRNAs in the lateral amygdala. *ELife*, *8*, e51607. https://doi.org/10.7554/eLife.51607

Pérez de Sevilla Müller, L., Do, M. T. H., Yau, K.-W., He, S., & Baldridge, W. H. (2010). Tracer coupling of intrinsically photosensitive retinal ganglion cells to amacrine cells in the mouse retina. *The Journal of Comparative Neurology*, *518*(23), 4813–4824. https://doi.org/10.1002/cne.22490

Pfeiffer, B. E., & Huber, K. M. (2006). Current Advances in Local Protein Synthesis and Synaptic Plasticity. *Journal of Neuroscience*, *26*(27), 7147–7150. https://doi.org/10.1523/JNEUROSCI.1797-06.2006

Preußner, M., & Heyd, F. (2016). Post-transcriptional control of the mammalian

circadian clock: Implications for health and disease. *Pflügers Archiv - European Journal of Physiology*, *468*(6), 983–991. https://doi.org/10.1007/s00424-016-1820-y

Pons, P. & Latapy, M. (2005). Computing Communities in Large Networks using Random Walks. Heidelberg Berlin: Springer.

Provencio, I., Jiang, G., De Grip, W. J., Hayes, W. P., & Rollag, M. D. (1998). Melanopsin: An opsin in melanophores, brain, and eye. *Proceedings of the National Academy of Sciences*, *95*(1), 340–345. https://doi.org/10.1073/pnas.95.1.340

Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E., & Zanini, F. (2022). Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics*, *38*(10), 2943–2945. https://doi.org/10.1093/bioinformatics/btac166

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rao, S., Chun, C., Fan, J., Kofron, J. M., Yang, M. B., Hegde, R. S., Ferrara, N., Copenhagen, D. R., & Lang, R. A. (2013). A direct and melanopsin-dependent fetal light response regulates mouse eye development. *Nature*, *494*(7436), 243–246. https://doi.org/10.1038/nature11823

Renna, J. M., Weng, S., & Berson, D. M. (2011). Light acts through melanopsin to alter retinal waves and segregation of retinogeniculate afferents. *Nature Neuroscience*, *14*(7), 827–829. https://doi.org/10.1038/nn.2845

Reynoso, M. A., Juntawong, P., Lancia, M., Blanco, F. A., Bailey-Serres, J., & Zanetti, M. E. (2015). Translating Ribosome Affinity Purification (TRAP)

Followed by RNA Sequencing Technology (TRAP-SEQ) for Quantitative Assessment of Plant Translatomes. In J. M. Alonso & A. N. Stepanova (Eds.), *Plant Functional Genomics* (Vol. 1284, pp. 185–207). Springer New York. https://doi.org/10.1007/978-1-4939-2444-8_9

Risso, D., Ngai, J., Speed, T. P., & Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, *32*(9), 896–902. https://doi.org/10.1038/nbt.2931

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47–e47. https://doi.org/10.1093/nar/gkv007

Rozenbaum, M., Rajman, M., Rishal, I., Koppel, I., Koley, S., Medzihradszky, K. F., Oses-Prieto, J. A., Kawaguchi, R., Amieux, P. S., Burlingame, A. L., Coppola, G., & Fainzilber, M. (2018). Translatome Regulation in Neuronal Injury and Axon Regrowth. *Eneuro*, *5*(2), ENEURO.0276-17.2018. https://doi.org/10.1523/ENEURO.0276-17.2018

Sadiq, M. A., Hanout, M., Sarwar, S., Hassan, M., Agarwal, A., Sepah, Y. J., Do, D. V., & Nguyen, Q. D. (2016). Platelet-Derived Growth Factor Inhibitors: A Potential Therapeutic Approach for Ocular Neovascularization. In Q. D. Nguyen, E. B. Rodrigues, M. E. Farah, W. F. Mieler, & D. V. Do (Eds.), *Developments in Ophthalmology* (Vol. 55, pp. 310–316). S. Karger AG. https://doi.org/10.1159/000438953

Sanz, E., Yang, L., Su, T., Morris, D. R., McKnight, G. S., & Amieux, P. S. (2009).

Cell-type-specific isolation of ribosome-associated mRNA from complex tissues. *Proceedings of the National Academy of Sciences*, *106*(33), 13939–13944. https://doi.org/10.1073/pnas.0907143106

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., … Church, D. M. (2016). *Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly* [Preprint]. Genomics. https://doi.org/10.1101/072116

Sernagor, E. (2005). Retinal Development: Second Sight Comes First. *Current Biology*, *15*(14), R556–R559. https://doi.org/10.1016/j.cub.2005.07.004

Sheerin, D., O'Connor, D., Pollard, A. J., & Mohorianu, I. (2019). *Effects of technical noise on bulk RNA-seq differential gene expression inference* [Preprint]. Bioinformatics. https://doi.org/10.1101/843789

Shigeoka, T., Jung, H., Jung, J., Turner-Bridger, B., Ohk, J., Lin, J. Q., Amieux, P. S., & Holt, C. E. (2016). Dynamic Axonal Translation in Developing and Mature Visual Circuits. *Cell*, *166*(1), 181–192. https://doi.org/10.1016/j.cell.2016.05.029

Shigeoka, T., Jung, J., Holt, C. E., & Jung, H. (2018). Axon-TRAP-RiboTag: Affinity Purification of Translated mRNAs from Neuronal Axons in Mouse In Vivo. In I. Gaspar (Ed.), *RNA Detection* (Vol. 1649, pp. 85–94). Springer New York. https://doi.org/10.1007/978-1-4939-7213-5_5

Sudhesh Dev, S., Zainal Abidin, S. A., Farghadani, R., Othman, I., & Naidu, R. (2021). Receptor Tyrosine Kinases and Their Signaling Pathways as Therapeutic Targets of Curcumin in Cancer. *Frontiers in Pharmacology, 12*, 772510. https://doi.org/10.3389/fphar.2021.772510

Sutton, M. A., Wall, N. R., Aakalu, G. N., & Schuman, E. M. (2004). Regulation of Dendritic Protein Synthesis by Miniature Synaptic Events. *Science, 304*(5679), 1979–1983. https://doi.org/10.1126/science.1096202

Tian, L., Dong, X., Freytag, S., Lê Cao, K.-A., Su, S., JalalAbadi, A., Amann-Zalcenstein, D., Weber, T. S., Seidi, A., Jabbari, J. S., Naik, S. H., & Ritchie, M. E. (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature Methods*, *16*(6), 479–487. https://doi.org/10.1038/s41592-019-0425-8

Tiriac, A., Smith, B. E., & Feller, M. B. (2018). Light Prior to Eye Opening Promotes Retinal Waves and Eye-Specific Segregation. *Neuron*, *100*(5), 1059-1065.e4. https://doi.org/10.1016/j.neuron.2018.10.011

Townes, F. W., Hicks, S. C., Aryee, M. J., & Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology, 20*(1), 295. https://doi.org/10.1186/s13059-019-1861-6

Tushev, G., Glock, C., Heumüller, M., Biever, A., Jovanovic, M., & Schuman, E. M. (2018). Alternative 3′ UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments. *Neuron*, *98*(3), 495-511.e6. https://doi.org/10.1016/j.neuron.2018.03.030

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Vihanto, M. M., Plock, J., Erni, D., Frey, B. M., Frey, F. J., & Huynh-Do, U. (2005). Hypoxia up-regulates expression of Eph receptors and ephrins in mouse skin. *The FASEB Journal*, *19*(12), 1689–1691. https://doi.org/10.1096/fj.04-3647fje

Welshhans, K., & Bassell, G. J. (2011). Netrin-1-Induced Local -Actin Synthesis and Growth Cone Guidance Requires Zipcode Binding Protein 1. *Journal of Neuroscience*, *31*(27), 9800–9813. https://doi.org/10.1523/JNEUROSCI.0166-11.2011

Yao, J., Sasaki, Y., Wen, Z., Bassell, G. J., & Zheng, J. Q. (2006). An essential role for β-actin mRNA localization and translation in Ca2+-dependent growth cone guidance. *Nature Neuroscience*, *9*(10), 1265–1273. https://doi.org/10.1038/nn1773

Zhang, D.-Q., Wong, K. Y., Sollars, P. J., Berson, D. M., Pickard, G. E., & McMahon, D. G. (2008). Intraretinal signaling by ganglion cell photoreceptors to dopaminergic amacrine neurons. *Proceedings of the National Academy of Sciences*, *105*(37), 14181–14186. https://doi.org/10.1073/pnas.0803893105

Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). ComBat-seq: Batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics*, *2*(3), lqaa078. https://doi.org/10.1093/nargab/lqaa078

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T.,

Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*(1), 14049. https://doi.org/10.1038/ncomms14049