Spring 2021

Transcriptional Profiling of Neurological Development of Drosophila Following Bisphenol A Exposure

Eden Johnson
San Jose State University

Transcriptional Profiling of Neurological Development of Drosophila Following Bisphenol A Exposure

A Project

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfillment

Of the Requirements for the

Degree Master of Science in Bioinformatics

By

Eden Johnson

May 2021

ABSTRACT

The ubiquitous environmental chemical bisphenol A (BPA) is an emerging risk factor for neurodevelopmental disorders (NDDs). BPA is an endocrine-disrupting chemical (EDC) that is thought to interfere with neuron development by changing neuronal gene expression. Impacting neurodevelopment in this manner can cause lasting changes in behavior and potentially lead to the development of NDDs. Delineating the molecular processes that incur changes in gene expression following BPA exposure will advance our understanding of how BPA impacts neurodevelopmental pathways and affects the pathophysiology of NDDs. An RNA-Sequencing (RNA-Seq) analysis pipeline was created for transcriptional profiling of neurological development in wild-type *Drosophila melanogaster* following BPA exposure. The analysis pipeline identified differentially expressed genes and the affected molecular pathways. The results indicated changes gene expression in neurodevelopmentally-related pathways in the wild-type background.

Index terms - Bisphenol A (BPA), neurodevelopmental disorders (NDDs), endocrine-disrupting chemical (EDC), transcriptional profiling, RNA-Sequencing (RNA-Seq)

ACKNOWLEDGEMENTS

I wanted to first and foremost thank Dr. Wendy Lee for her constant guidance and mentorship throughout this project and the MS in Bioinformatics program. Her advice has helped me not only academically, but professionally and personally. Her insight helped keep me motivated and productive throughout our collaboration. I am thankful to have gotten the opportunity to work with her.

I want to recognize Dr. Mulligan for serving as a committee member and for her collaboration on this project. Dr. Mulligan provided invaluable insight, feedback, and time on this project and for that I am extremely appreciative and thankful. This project would not have been possible without her passion for this area of research.

I want to recognize Dr. Heller for serving as a committee member as well and for providing his support and feedback regarding this project. Dr. Heller has also been a wonderful instructor in the MS in Bioinformatics program and I want to thank him for taking the time to be a member on this committee.

I want to lastly thank Story and my family; thank you for always supporting me through the long hours and keeping me motivated. Thank you for always believing in my work and for helping me stay grounded throughout this process.

TABLE OF CONTENTS

LIST OF FIGURES

List Of Tables

1. Introduction

1.1 Biological Background

The prevalent environmental chemical bisphenol A (BPA)—used in the synthesis of plastics and found in a vast number of common consumer products—can interfere with neurodevelopmental pathways and negatively affect the formation of neurons [1], [2]. The lipophilic structure of BPA allows it to easily cross the placental barrier to potentially harm a developing fetus [1]; in fact, maternal BPA has been detected in the placenta [3]. BPA can act as an estrogen mimic and is categorized as an endocrine disrupting chemical (EDC), although it likely interrupts signaling pathways beyond estrogenic and androgenic pathways to impact brain development. A variety of adverse neurodevelopmental consequences of prenatal exposure to BPA have been documented. For instance, in experiments performed by Arambula et al. 2016, prenatal BPA exposure in developing rats was shown to disrupt gene expression in the hypothalamus, even at low doses [1]. By influencing gene expression in the developing hypothalamus, the region of the brain that regulates innate social behavior, BPA exposure impacts transcription-level changes that can result in behavioral changes [1]. A study conducted by Tiwari et al. 2014 found that prenatal BPA exposure in rats down-regulated genes involved with myelination of axons in the hippocampus, resulting in decreased hippocampal myelination and impaired cognitive function [4]. Other studies conducted in the field have indicated BPA can impact a battery of neurodevelopmentally-relevant cellular phenotypes, including neural stem cell speciation, proliferation, and axon guidance [5], [6]. Due to its environmental prevalence and increasing evidence indicating its ability to disrupt development, there is rightful concern regarding BPA exposure and its suspected contribution to neurotoxicity.

The emergent data indicating the developmental neurotoxicity of BPA has led to alarm surrounding BPA as an environmental risk factor for neurodevelopmental disorders (NDDs). Many NDDs, including autism spectrum disorders (ASD), have both environmental and genetic etiologies [7]**.** The "gene by environment hypothesis" posits that environmental factors often work in concert with genetic risk factors to confer the greatest risk of NDDs; thus, BPA potentially operates in association with genetic predispositions for NDDs to hinder neural development [7]. A study conducted by Stein et al. 2016 revealed an association between higher levels of BPA and children with ASD [7]. The increased frequency in NDD diagnoses in recent years, most notably ASD, is of urgent concern; the increased prevalence of both BPA in common materials and evidence supporting a role in neurodevelopmental impairment warrants investigation [8]. Despite the mounting evidence that BPA can impair neural development, the specific molecular mechanisms affected are not well characterized [9]. The discovery of the characterizations of these mechanisms will help advance drug development to treat NDDs and could help inform new social policy to reduce environmental BPA.

The common fruit fly, *Drosophila melanogaster* (*D. melanogaster*), has long been used as a model organism for human diseases and for developmental biology research. In recent years, *D. melanogaster* has increasingly been used for investigations in toxicology given its low cost, short life cycle, and ability to be easily maintained and exposed to environmental chemicals [10]. In the laboratory of Dr. Kimberly Mulligan at California State University, Sacramento, wild-type *D. melanogaster* larvae were exposed to BPA in order to investigate the neurodevelopmental impacts at both the cellular and behavioral levels [6]. Using flies for this study allows disentangling EDC-related vs non-EDC-related neurodevelopmental impacts, since flies lack

estrogen receptors and hence do not have the confounding factor of EDC-related impacts. In recently published work by Mulligan's research group, BPA exposure was shown to impact axon guidance, neuroblast development, locomotor behavior, grooming activity, and courtship behaviors in wild-type *D. melanogaster* [6]. Further, axon guidance was specifically impaired in the mushroom body, an adult neural structure required for olfactory-based learning and memory [6]. Indeed, unpublished research from the Mulligan research group indicates that developmental BPA exposure also affects learning memory in adult *D. melanogaster* (personal communication). However, the molecular mechanisms influencing these cellular and behavioral changes are unknown. Using RNA sequencing analysis to identify transcriptional changes caused by BPA will help elucidate the specific molecular changes that underlie these neurodevelopmental outcomes.

The aim of this project was to establish an RNA sequencing pipeline that can be used for investigating how BPA influences neurodevelopmentally-related pathways in *D. melanogaster*. In addition, this pipeline was established to validate and provide molecular explanations for the cellular and behavioral outcomes observed in the laboratory of Dr. Kimberly Mulligan. This project proposes the incorporation of RNA sequencing (RNA-Seq) analysis through the use of a pipeline that runs on the samples generated by next-generation sequencing. The output of the pipeline is a set of genes found to be statistically significant in terms of differential expression among the two treatment conditions—wild-type flies exposed to BPA exposed compared to unexposed wild-type control flies. In addition, the RNA-Seq pipeline identifies neurodevelopmentally-relevant molecular pathways that the genes of interest are involved in. The overall goal of the pipeline was to identify genes and molecular pathways of interest that

have neurodevelopmental significance in order to determine the transcriptional impact of BPA as it pertains to neurodevelopment.

This transcriptional profiling using RNA-Seq analysis offers deeper insight into which neurodevelopmentally-related molecular processes are being affected by BPA exposure. RNA-Seq analysis has been used in practice to perform differential expression analysis and to investigate the change of expression between treatment groups [11]. The benefits of using RNA-Seq for transcriptional profiling in this particular application is that it will comprehensively delineate molecular pathways of neurological significance and help characterize specific genes that may potentially serve as targets in drug development for the treatment of NDDs. The RNA-Seq pipeline incorporated the use of the Bioinformatics tools FastQC, Trimmomatic, HiSAT2, HTSeq-Count, DESeq2, and GSEA to have an end-to-end workflow.

1.2 Technical Background

The general workflow of the pipeline analysis includes a preprocessing step that includes a quality check and quality trimming, read mapping and genome indexing, the marking and removal of duplicates, mapped transcript quantification, differential expression analysis, and gene set enrichment analysis.

1.2.1 NGS Read Quality Control

FastQC was used in the preprocessing step of the workflow to evaluate the quality of the sequenced reads [12]. FastQC evaluates raw sequencing data and marks possible problem areas that can hopefully be resolved with other preprocessing tools. FastQC provides metrics that show

the quality of the raw sequencing data based upon per base sequence quality, per sequence quality scores, sequence duplication levels, overrepresented sequences, and adapter content [12]. FastQC is used before and after other preprocessing tools to ensure high data quality prior to data analysis. FastQC was chosen because it is an industry standard for read quality evaluation.

1.2.2 Sequencing Adapter and Quality Trimming

Trimmomatic is a tool that was incorporated in the pre-processing step of the RNA-sequencing workflow and works to remove any adapters or primers that were used during the preparation step for sequencing [13]. Trimmomatic works by taking in the raw sequencing data files generated from sequencing and performs read trimming and filtering to result in high quality and relevant reads [13]. The filtering step is based upon a user-input threshold, in this case a Phred quality score of 33. A local alignment is performed between the adapters and reads; reads with below-threshold phred quality scores are removed [13]. This filtering step is essential for maintaining the integrity of the sequenced reads for the downstream analysis. Trimmomatic was chosen for quality filtering and trimming of adapters as Illumina adapters were used for sequencing.

1.2.3 Read Mapping and Genome Indexing

HiSAT2 (hierarchical indexing for spliced alignment of transcripts 2) is an alignment that can be used for next-generation sequencing read-mapping of RNA when provided a reference genome [14]. When using HiSAT2 with a reference genome for read mapping, two major tasks are performed—genome indexing and read mapping. In the genome indexing step, HiSAT2 works to index a reference genome by utilizing a graph-based indexing. When provided with a

genome-annotation file (GTF) that contains the splice site information of the genome, this indexing scheme works to increase the speed of read mapping in the next step [14]. The mapping step works by using the HiSAT2 created-indexes and a graph-based alignment approach to result in a reliable alignment [14]. With adjustable parameters for minimum and maximum mismatch penalties as well as its speed, HiSAT2 was chosen as the alignment tool for the RNA sequencing pipeline. The HiSAT2 protocol proposed by Pertea et al. 2016 was incorporated into the workflow [15].

Samtools contains a utility toolset that can manipulate the SAM file format [16]. The Samtools toolset includes the option to convert .sam files into .bam files. The BAM files, which are compressed, are the binary formatted version of SAM files.

1.2.4 Mark Duplicates

Picard contains command-line tools for manipulating data in the SAM/BAM format [17]. One tool that was utilized was the MarkDuplicates feature, which takes in reads in the SAM/ BAM format and marks them as duplicates and then outputs them in the BAM format, with duplicates marked or removed when provided the appropriate flag [17]. The MarkDuplicates tool was utilized in this pipeline to account for the possibility of PCR duplicates and to remove them before the count step to further validate results generated by the count files.

1.2.5 Transcript Quantification

HTSeq (high-throughput sequencing) count is a tool that can count how many reads map to certain genomic features, given a file with aligned reads in the SAM/BAM format and a genome annotation file with features to reference [18]. This helps quantify how many of the sequenced

reads align to known transcripts and genomic features. The count files generated by HTSeq count serve as input for downstream differential expression analysis.

1.2.6 Differential Expression Analysis

DESeq2 performs differential expression analysis on count data [11]. When performing RNA-Seq analysis, a desired result is the identification of genes that are being differentially expressed across different sample conditions [11]. DESeq2 creates a dispersion estimation using the negative binomial (Gamma-Poisson) distribution, which results in assigning the input genes p-values, p-adj values, and log2 fold change values between control and treatment groups. The DE analysis step results in the identification of genes that are differentially expressed between the sample conditions. The results can be further refined by filtering the results table to only include genes with p-adj values < 0.01. This significance level threshold can further validate that identified genes were not differentially expressed due to chance, and show a notable difference in expression between the control and treatment samples. A main metric for comparison is the log2 fold change (LFC) between two different treatment groups; a LFC value of zero would indicate no differential expression between treatment conditions for a particular gene [11].

1.2.7 Gene Set Enrichment Analysis

Fgsea (fast gene set enrichment analysis) is a library in R that performs fast gene set enrichment analysis on a set of ranked genes based upon a metric [19]. The metrics used can be p-values, p-adj values, log2 fold changes, or can be obtained from a pre-ranked list. The algorithm used by fgsea results in statistically significant p-values associated to molecular pathways and genes within those pathways. The Molecular Signatures Database (MSigDB) is

often used as a reference database when performing gene set enrichment analysis as it contains common pathways and associated genes specific to model organisms, such as *D. melanogaster* [20]–[23]**.** The ranked gene list provided as input can be compared to the *D. melanogaster* pathways and genes identified by MSigDB; this comparison can be used to identify the significant genes and corresponding pathways that are up-regulated and down-regulated across sample conditions. Delineating the pathways impacted by differential genes can provide critical insight into the specific molecular networks that are disrupted following exposure to an environmental chemical such as BPA. GSEA can be performed directly through the Broad institute application; however, fgsea in R was chosen to minimize the use of additional software dependencies.

1.2.8 Pipeline Execution Tool

Snakemake is a Python-based workflow that works to deploy multiple tools in a variety of environments, including high performance computing (HPC), cluster, and cloud environments [24]. The Snakemake workflow takes in expected input and output and runs tools based upon "rules" in a Snakefile, similar to a script. Snakemake makes highly scalable data analysis workflows that allow for ease of execution of popular Bioinformatics analysis tools that have their own Snakemake wrappers [24]. Snakemake was utilized in the execution of this RNA-Seq pipeline given its compatibility with the various Bioinformatics tools used in the analysis, and its reproducibility in multiple types of environments.

2. Methods

2.1 Tools/Pipeline Execution

The sequencing data was generated from four biological replicates, four samples referred to as the wild-type condition (unexposed wild-type flies), and four samples referred to as the BPA-exposed condition (wild-type flies exposed to 1 mM of BPA throughout development). The samples were sequenced on the HiSeq 4000 platform at 100 bp reads to generate single end data (SR100) at the UC Davis genomics core. The sample names, their condition, and the relative yields can be found in Table I below.

Table I. Raw sequencing data reference

Sample Name	Condition	Yield (# of reads)
A_2_S123_L004_R1_001	1 mM BPA	37,602,008
A_S119_L004_R1_001	Wild-type (0 mM BPA)	35,233,735
B_2_S124_L004_R1_001	1 mM BPA	43,739,934
B_S120_L004_R1_001	Wild-type (0 mM BPA)	36,384,564
C_2_S125_L004_R1_001	1 mM BPA	37,941,942
C_S121_L004_R1_001	Wild-type (0 mM BPA)	38,523,284
D_2_S126_L004_R1_001	1 mM BPA	37,090,520
D_S122_L004_R1_001	Wild-type (0 mM BPA)	36,610,735

The overall workflow of this pipeline follows the flowchart seen in Figure 1 below. The flowchart includes the major steps of the pipeline denoted in bold, the tools incorporated in the

pipeline denoted in boxes, as well as input and output of each tool.

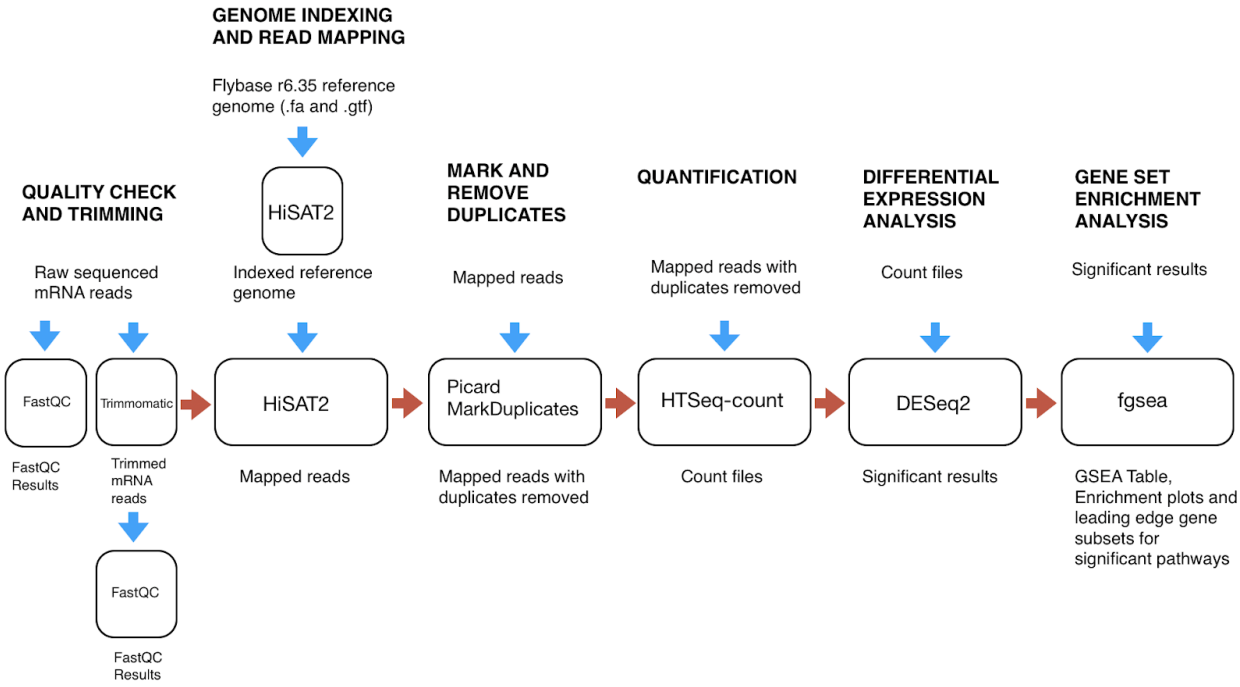

Figure 1. Pipeline Flowchart Overview

In the preprocessing step, FastQC (0.11.8) was used to evaluate the raw sequencing data quality in the HTML output files. The FastQC output was also used to identify which Illumina adapters needed to be trimmed using Trimmomatic (0.39). Trimmomatic was used to trim the TruSeq Adapter, Index 12 'GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGCCGTCTTCTGCTTG', and RNA PCR Primer Index 12 'CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTC' adapters from the raw sequencing data. Trimmomatic was used in the single-end read mode, as the raw sequencing data was not paired ended, and a phred quality score threshold of 33 was used to only have high-

quality reads after the trimming step. These trimmed reads were evaluated once again using FastQC to ensure proper trimming of adapters and to validate read quality and to ensure enough reads were still present after using the Trimmomatic tool to perform downstream analysis.

A reference genome fasta file and genome annotation file in the form of general transfer format (gtf) were obtained from Flybase, for the FB2020_04, dmel.r6.35 version [25]. HiSAT2 (2.1.0) was used for genome indexing provided the genome fasta file and the genome annotation file. Read mapping was performed in HiSAT2 using the newly indexed reference genome and the trimmed reads output by Trimmomatic. For HiSAT2, the `--mp` flag was adjusted to `4,2` the maximum and minimum penalties for mismatches. The read mapping step resulted in .sam files, which are a common format for storing aligned reads and their reference sequences. The SAM files were then sorted and converted into BAM files to get the files into binary format using Samtools (1.4.1). The Picard tool, MarkDuplicates (2.21.8), was used on the mapped BAM files output by Samtools, and duplicates were marked and removed and output into new BAM files. The types of duplicates that can be removed from the samples include sequencing duplicates and PCR duplicates.

For read feature quantification, HTSeq count (0.11.3) was used on the deduplicated mapped BAM files created by Picard. To deal with reads that align with more than one feature, HTSeq count provides three different modes, `union, intersection_strict,` and `intersection_nonempty` [18]. When not specified, the default mode is union, and the default mode was used in this pipeline as it is the mode recommended for most cases. HTSeq count was provided the reads with duplicates removed, and the genome annotation file (gtf) as input. The input genome annotation file (gtf) contained Ensembl annotated transcripts to

associate the reads with known transcripts in the reference genome [26]. Provided these inputs, HTSeq count generated count files containing counts for each known transcript per sample. Specific parameters that were set were the parameters `-s` to `no`, `-i` to `transcript_id` to obtain the Ensembl transcript ID for the mapped read, and `--additional-attr` to `gene_id`. These count files were used as input for differential expression (DE) analysis.

DE analysis was performed using the DESeq2 package (1.28.1) in R. DESeq2 works by using the count files generated by HTSeq count as input, providing sample names and each sample's corresponding condition (treatment or control), and converting the count files into statistically significant datasets, with specific genes identified as being more statistically significant than others based upon p-adj values at threshold of 0.01 in the current usage of this application. Since the input count files contained the *D. melanogaster* Ensembl transcript ID rather than the gene symbol the transcripts are associated with, the R library, biomaRt (2.44.4), was used to retrieve the entire Ensembl entry for *D. melanogaster* in order to get the needed gene symbols for the gene list. The table of significant genes with their corresponding gene symbols and statistics were output into a `.csv` file and ordered by p-adj value, in ascending order.

Gene set enrichment analysis was performed using the fgsea package (1.14.0) in R. The ranked gene list was created based upon the `.csv` file output from DESeq2 and contained only genes with associated gene symbols, and the log2 fold change metric was used as the ranking metric. This ranked gene list had duplicates removed, on the occasion that multiple transcripts observed in the results mapped to the same gene. The msigdb package was used to gather molecular pathways and their corresponding genes listed for *D. melanogaster* to demonstrate the significance of differentially expressed neurodevelopmentally related pathways when compared

across molecular processes from the Gene Ontology (GO) collection in the Cellular Component (CC) and Biological Processes (BP) subcategories. The fgsea algorithm was used after being provided the ranked gene list from DESeq2 results and the gene sets for *D. melanogaster* as input for the GO collection, CC and BP subcategories. As output, a gene set enrichment table was created, along with a text file containing the top 10 up-regulated and down-regulated pathways, their statistical significance, and a list of their leading edge genes. The leading edge genes represent the genes in a specific pathway with the highest contribution to the enrichment score. Genes that appear in multiple leading edge subsets for different pathways are likely subject to differential expression following a treatment vs. a control condition [20]–[23]. An enrichment plot was created for each of the top 10 up-regulated and top 10 down-regulated pathways observed in the gene set enrichment table.

2.2 Implementation of Pipeline Execution

The analysis pipeline was designed to perform an end-to-end workflow. The steps of Snakemake are split into 9 major steps for pipeline execution on the SJSU Spartan HPC (High Performance Computer):

- FastQC on the raw sequencing reads
- Trimmomatic on the raw sequencing reads
- FastQC on the trimmed reads
- HiSAT2 indexing
- HiSAT2 read mapping of the trimmed reads
- Removal of all types of duplicates from the mapped reads with Picard

- HTSeq-count for mapped read quantification
- Differential expression analysis on the count files using DESeq2
- Gene set enrichment analysis on the statistically significant differentially expressed genes using fgsea

The requirements for the pipeline are as follows:

- Linux system with scalable nodes
- FastQC (0.11.8)
- Trimmomatic (0.39)
- HiSAT2 (2.1.0)
- Samtools (1.4.1)
- Picard MarkDuplicates (2.21.8)
- Python 3+
- HTSeq Count (0.11.3)
- R 4.0+
- DESeq2 (1.28.1)
- fgsea (1.14.0)
- Snakemake (4.3.0)

The main pipeline call is in a single shell script that is intended to be deployed in an HPC environment. In the script, the pipeline was allocated 8 compute nodes and 8 CPUs for each node, and 10 GB per node. For a reference genome size of 180 million base pairs (bp) for *D.*

melanogaster, full pipeline-execution takes ~3.5 hours from start to finish when analyzing 8 samples with raw library sizes ranging from 35,000,000 to 45,000,000 [27]. Execution is dependent upon the library size of the trimmed samples, the number of samples, and the size of the reference genome. This pipeline is scalable to allocate more nodes and CPUs for processing. In addition, since the pipeline is deployed using Snakemake, it can easily be deployed in other environments.

3. Results

3.1 Read Quality Filtering and Trimming

The first round of FastQC results generated HTML files that contained graphs and visuals for each sample processed. The FastQC results indicated that Illumina Universal adapters were present in the raw sequencing, and therefore a trimming step would be required. The FastQC results for the raw sequencing reads can be found in the appendix.

Table II includes the results for each of the 8 samples with input, output, and dropped read counts, as well as percentages of reads surviving and reads dropped following trimming. As observed in Table II, the number of reads dropped per sample following trimming at a phred score threshold of 33 resulted in less than 1% of reads being dropped in each sample. Table II shows the number of reads before and after the trimming step, the number of reads dropped, as well as the percent of reads dropped and percent surviving for each sample following the trimming step.

Table II. Trimmomatic Results

Sample Name	Raw Input Reads	Surviving	% Surviving	Dropped	% Dropped
A_2_S123_L004_R1_001	37,602,008	37,521,543	99.79%	80,465	0.21%
A_S119_L004_R1_001	35,233,735	35,160,114	99.79%	73,621	0.21%
B_2_S124_L004_R1_001	43,739,934	43,655,424	99.81%	84,510	0.19%
B_S120_L004_R1_001	36,384,564	36,312,251	99.80%	72,313	0.20%
C_2_S125_L004_R1_001	37,941,942	37,860,379	99.79%	81,563	0.21%
C_S121_L004_R1_001	38,523,284	38,460,091	99.84%	61,193	0.16%
D_2_S126_L004_R1_001	37,090,520	37,013,085	99.79%	77,435	0.21%
D_S122_L004_R1_001	36,610,735	36,547,791	99.83%	62,944	0.17%

A second FastQC check was performed on the trimmed reads to ensure proper removal of the adapters from the raw reads, and to ensure only high quality reads remained after the trimming step. FastQC results were observed in the form of an HTML file to confirm that analysis could proceed with passing values for each FastQC metric. The FastQC results from the trimmed reads can be found in the appendix. The trimmed read files served as input files for the read mapping step using HiSAT2.

3.2 Read Mapping and Removal of Duplicates

HiSAT2 created an indexed reference genome from the dmel-all-r6.35 reference genome and genome annotation file (gtf) obtained from Flybase. The indexes were used by HiSAT2 to

map the reads. The mapping rates for each of the 8 samples can be observed in Table III. Table III includes metrics to show the percentage of reads that aligned 0, 1, or greater than 1 time and the final overall alignment percentage. All samples processed through the pipeline had alignment rates higher than 78% when mapped to the indexed reference genome. In addition, less than 25% of the overall reads for each sample did not align to the indexed reference genome at all. All mapped reads were used for downstream analysis.

Table III. HiSAT2 Results

Sample Name	Trimmed Input reads	% Aligned 0 times	% Aligned exactly 1 time	% Aligned >1 times	Overall alignment rate
A_2_S123_L004_R1_001	37,521,543	14.41%	83.79%	1.79%	**85.59%**
A_S119_L004_R1_001	35,160,114	21.82%	77.06%	1.11%	**78.18%**
B_2_S124_L004_R1_001	43,655,424	16.56%	81.93%	1.51%	**83.44%**
B_S120_L004_R1_001	36,312,251	19.12%	79.18%	1.70%	**80.88%**
C_2_S125_L004_R1_001	37,860,379	17.86%	80.36%	1.78%	**82.14%**
C_S121_L004_R1_001	38,460,091	19.50%	78.03%	2.46%	**80.50%**
D_2_S126_L004_R1_001	37,013,085	16.66%	82.13%	1.21%	**83.34%**
D_S122_L004_R1_001	36,547,791	21.90%	76.65%	1.46%	**78.10%**

The mapped reads were then processed by Samtools to produce BAM files to get the mapped read files into binary format for usage by other tools. The newly created mapped BAM files were then processed by the Picard tool MarkDuplicates. MarkDuplicates was run twice during the testing portion of the pipeline for comparing counts and percentages of surviving reads compared to raw input reads; once with the optional flag of `REMOVE_DUPLICATES` set to `true` to remove any type of duplicate from the file, and another time with the optional flag

`REMOVE_SEQUENCING_DUPLICATES` set to `true` to remove optical or sequencing duplicates from the file. The counts and percentages of reads remaining for each type of duplicate removal compared to the raw input reads in each of the 8 samples in comparison to each raw input read counts can be found in Table IV.

Table IV. MarkDuplicates Results

Sample Name	Raw Input Reads	Reads Remaining: Remove Sequencing Duplicates	Reads Remaining: Remove Sequencing Duplicates (Percentage)	Reads Remaining: Remove All Duplicates	Reads Remaining: Remove All Duplicates (Percentage)
A_2_S123_L004_R1_001	37,602,008	34,341,679	91.33%	3,143,138	8.36%
A_S119_L004_R1_001	35,233,735	29,017,175	82.36%	1,743,988	4.95%
B_2_S124_L004_R1_001	43,739,934	38,813,870	88.74%	3,371,633	7.71%
B_S120_L004_R1_001	36,384,564	31,542,053	86.69%	2,475,934	6.81%
C_2_S125_L004_R1_001	37,941,942	33,987,880	89.58%	4,021,337	10.60%
C_S121_L004_R1_001	38,523,284	33,703,298	87.49%	3,103,708	8.06%
D_2_S126_L004_R1_001	37,090,520	32,624,585	87.96%	2,305,966	6.22%
D_S122_L004_R1_001	36,610,735	30,664,167	83.76%	2,344,468	6.40%

3.3 Counting

HTSeq Count was used to create count files based upon the mapped reads with duplicates removed. The count files include information about the Ensembl gene ID and the Ensembl transcript ID the reads mapped to and the counts of how many reads were mapped to these known transcript entries. These count files served as input for differential expression analysis with DESeq2.

3.4 Differential Expression Analysis

DESeq2 was used in R to perform differential expression analysis. DESeq2 produced a table of significant genes, with p-values, p-adj values, and log2 fold changes as some noted associated statistics. A threshold of a p-adj value of 0.01 was used to filter out genes that were not found to be statistically significant. The significant genes were output into a `.csv` file and ordered by p-adj value, in ascending order and were used as input to perform gene set enrichment analysis. The DESeq function of DESeq2 runs a diagnostic test for outliers, called *Cook's distance* for every gene and for every sample. Cook's distance is a measure of the influence an individual sample has over the fitted coefficients for a particular gene; a large Cook's distance value can be used for inferring a high outlier count [11]. Furthermore, genes with low counts are due to high dispersion [11]. Table V below includes a summary of the significant results, including the counts and percentages of genes that experienced a positive log2 fold change, a negative log2 fold change, as well as outliers and low counts.

Table V. Summary of Significant DESeq2 Results

Feature	Counts	Percentage
LFC > 0 (up)	597	57%
LFC < 0 (down)	443	43%
Outliers	0	0%
Low counts	0	0%
Total	**1040**	**100%**

3.5 Gene Set Enrichment Analysis (GSEA)

GSEA was used to help identify which processes and molecular pathways experienced a large log2 fold change between unexposed control samples and BPA-exposed samples. GSEA was used on the GO collection, with the Cellular Component (CC) and Biological Processes (BP) subcategories of *D. melanogaster* gene sets available in order to demonstrate the influence of BPA exposure on a focused set of neurodevelopmentally-relevant pathways. GSEA was performed using the fgsea library in R, which resulted in a table with the top 10 up-regulated and top 10 down-regulated pathways for both the GO CC and GO BP collections. The pathways in each collection were given an associated p-value, p-adjusted value, a NES (Normalized enrichment score), and an Enrichment Score (ES) to display statistical significance [20]–[23]. The p-value metric is an enrichment p-value, estimating the statistical significance of the ES for a specific gene set, and the p-adjusted value is a Benjamini-Hocheberg (BH)-adjusted p-value [20]–[23]. The ES denotes to which degree a gene set is overrepresented at the top (with a positive value) or at the bottom (with a negative value) of a ranked list of genes. The NES is the normalized ES that accounts for differences in gene set sizes by taking the actual ES and dividing by the mean of ESs against all permutations of the dataset, making NES comparable across gene sets [20]–[23]. Up-regulated pathways are associated with NES values greater than 0, and down-regulated pathways are associated with NES values less than 0. Tables VI and VII organize the qualifying pathways based upon descending NES.

Table VI. GSEA top up-regulated and down-regulated pathways for GO Collection, CC Subset

pathway	pval	padj	ES	NES
GO_SYNAPSE	6.4156E-07	0.00034	0.4994	2.6477
GO_NUCLEAR_BODY	1.8008E-06	0.00047	0.5008	2.6085
GO_PRESYNAPSE	5.7737E-06	0.0010	0.6826	2.4786
GO_CHROMATIN	8.2669E-05	0.0087	0.4450	2.2785
GO_NUCLEAR_CHROMOSOME	8.0105E-05	0.0087	0.4450	2.2770
GO_CELL_CORTEX	0.0005864	0.0343	0.6528	2.1293
GO_CELL_PROJECTION_MEMBRANE	0.0005045	0.0343	0.5540	2.1295
GO_NUCLEAR_SPECK	0.0005300	0.0343	0.4722	2.1430
GO_POSTSYNAPSE	0.0005400	0.0343	0.5042	2.1222
GO_GLUTAMATERGIC_SYNAPSE	0.0006525	0.0343	0.5391	2.1142
GO_DENDRITE_TERMINUS	0.0541082	0.2886	-0.9745	-1.3065
GO_CILIARY_TIP	0.0541082	0.2886	-0.9745	-1.3065
GO_RADIAL_SPOKE	0.0471312	0.2695	-0.7655	-1.4943
GO_SPINDLE_POLE	0.0437376	0.2654	-0.7077	-1.5560
GO_MOTILE_CILIUM	0.0408560	0.2589	-0.4888	-1.6005
GO_XY_BODY	0.0300601	0.2227	-0.9883	-1.3249
GO_DEUTEROSOME	0.0300601	0.2227	-0.9883	-1.3249
GO_TIM23_MITOCHONDRIAL_IMPORT_ INNER_MEMBRANE_TRANSLOCASE_COMPLEX	0.0155024	0.1751	-0.9147	-1.5330
GO_TERTIARY_GRANULE_LUMEN	0.0117841	0.1512	-0.9922	-1.3301
GO_CILIARY_PLASM	0.0070265	0.1067	-0.5559	-1.8897

Table VII. GSEA top up-regulated and down-regulated pathways for GO Collection, BP Subset

pathway	pval	padj	ES	NES
GO_NERVOUS_SYSTEM_PROCESS	4.5632E-06	0.0105	0.5820	2.6134
GO_POSITIVE_REGULATION_OF_RNA_ BIOSYNTHETIC_PROCESS	6.5687E-06	0.0105	0.4319	2.4935
GO_POSITIVE_REGULATION_OF_ TRANSCRIPTION_BY_RNA_POLYMERASE_II	9.5165E-06	0.0105	0.4579	2.4321
GO_REGULATION_OF_TRANSPORT	7.3319E-06	0.0105	0.4565	2.4896
GO_BIOLOGICAL_ADHESION	3.9518E-05	0.0158	0.4649	2.3600
GO_NEGATIVE_REGULATION_OF_MOLECULAR_ FUNCTION	4.0231E-05	0.0158	0.4237	2.3304
GO_NEGATIVE_REGULATION_OF_ CELLULAR_BIOSYNTHETIC_PROCESS	2.7439E-05	0.0158	0.3916	2.2941
GO_NEGATIVE_REGULATION_OF_NUCLEOBASE_ CONTAINING_COMPOUND_METABOLIC_PROCESS	2.35995E-05	0.0158	0.3870	2.2975
GO_REGULATION_OF_CELLULAR_LOCALIZATION	3.9927E-05	0.0158	0.5110	2.4245
GO_REGULATION_OF_CELL_POPULATION_ PROLIFERATION	3.6673E-05	0.0158	0.3816	2.2144
GO_PYRIMIDINE_CONTAINING_COMPOUND_ CATABOLIC_PROCESS	0.0148325	0.2078	-0.9922	-1.3263
GO_PYRIMIDINE_CONTAINING_COMPOUND_ BIOSYNTHETIC_PROCESS	0.0148325	0.2078	-0.9922	-1.3263
GO_NUCLEOSIDE_SALVAGE	0.0148325	0.2078	-0.9922	-1.3263
GO_NUCLEOSIDE_CATABOLIC_PROCESS	0.0148325	0.2078	-0.9922	-1.3263
GO_NUCLEOBASE_METABOLIC_PROCESS	0.0148325	0.2078	-0.9922	-1.3263
GO_NUCLEOBASE_CONTAINING_SMALL_ MOLECULE_CATABOLIC_PROCESS	0.0148325	0.2078	-0.9922	-1.3263
GO_NUCLEOBASE_CONTAINING_SMALL_ MOLECULE_BIOSYNTHETIC_PROCESS	0.0148325	0.2078	-0.9922	-1.3263
GO_GLYCOSOL_COMPOUND_CATABOLIC_PROCESS	0.0148325	0.2078	-0.9922	-1.3263
GO_GLYCOSOL_BIOSYNTHETIC_ CATABOLIC_PROCESS	0.0148325	0.2078	-0.9922	-1.3263
GO_CELLULAR_METABOLIC_COMPOUND_SALVAGE	0.0148325	0.2078	-0.9922	-1.3263

The subset of leading edge genes for each of the top up-regulated and down-regulated pathways in the GO CC and GO BP collections can be observed in Tables VIII and IX below. Tables VIII and IX also include a column “size” which denotes the size of the particular pathway after removing genes that were not present in the pre-ranked gene list [20]–[23].

Table VIII. GSEA pathways leading edge gene subsets for GO Collection, CC Subcategory

pathway	size	leadingEdge
GO_SYNAPSE	45	tty Nlg2 SK dlg1 ck Nlg4 Rtnl1 Adar milt CtBP CG5059 scrib CASK sgg 5-HT7 mbt fz4 CG9328 IRSp53 Nufip dsh Galphai Sh LanA rdgA Pp1-87B
GO_NUCLEAR_BODY	42	tou mle CG5059 cnc Atx2 Gmap nop5 pan Sh3beta dom CG12877 CG3335 CG6843 Sirt1 sima CG9328 CkIalpha dsh Cwc25 CG1234 ncm rdgA Pp1-87B Mi-2 Ino80 Usp7 CG1677 Pabp2 Rbp1-like nonA peb Srpk79D Slu7 Lk6 pico MYPT-75D
GO_PRESYNAPSE	15	Nlg2 dlg1 Nlg4 milt CtBP scrib CASK IRSp53 Nufip dsh Sh rdgA Pp1-87B
GO_CHROMATIN	40	tou Lim3 mle Hnf4 jing Pdp1 cnc Dp Acf Caf1-180 pan dom Ssrp CG7154 Sirt1 sima pros Nufip dsh Mi-2 Ino80 CG33051 salr His3.3B Lim1 emc nonA Srpk79D CkIIbeta Mrtf Hr4 CG7137 Smox sqd XRCC1
GO_NUCLEAR_CHROMOSOME	38	Lim3 mle Hnf4 jing Pdp1 cnc Dp Acf Caf1-180 pan dom Ssrp CG7154 Sirt1 sima pros Nufip dsh Pp1-87B Mi-2 Ino80 SMC2 CG33051 His3.3B Lim1 emc Mrtf Hr4 fs(1)h tna Smox sqd XRCC1
GO_CELL_CORTEX	11	dlg1 ck milt CtBP CASK Fim capu chb Galphai
GO_CELL_PROJECTION_MEMBRANE	17	aru tty dlg1 CASK Fim chb mgl CG9328 Kank Galphai Sh
GO_NUCLEAR_SPECK	28	tou CG5059 Atx2 Gmap dom CG3335 CG6843 sima CG9328 CkIalpha Cwc25 CG1234 ncm rdgA Pp1-87B Mi-2 CG1677 Pabp2 Rbp1-like nonA peb Srpk79D Slu7 MYPT-75D
GO_POSTSYNAPSE	23	Nlg2 SK dlg1 Nlg4 Rtnl1 milt CG5059 scrib sgg IRSp53 dsh Sh rdgA Pp1-87B zip Prosalpha4
GO_GLUTAMATERGIC_SYNAPSE	18	Nlg2 dlg1 Nlg4 CtBP scrib sgg fz4 IRSp53 dsh Sh rdgA Pp1-87B
GO_DENDRITE_TERMINUS	1	IFT57
GO_CILIARY_TIP	1	IFT57
GO_RADIAL_SPOKE	3	CG15143 CG15144 CG13436
GO_SPINDLE_POLE	4	SAK fzy asp

GO_MOTILE_CILIUM	11	CG15143 CG10252 CG12020 CG15144 CG13436 CG16719 CG10064
GO_XY_BODY	1	SAK
GO_DEUTEROSOME	1	SAK
GO_TIM23_MITOCHONDRIAL_IMPORT_ INNER_MEMBRANE_TRANSLOCASE_ COMPLEX	2	ttm3 CG7382
GO_TERTIARY_GRANULE_LUMEN	1	CG8349
GO_CILIARY_PLASM	12	CG6652 IFT57 CG15143 CG12020 CG15144 CG13436 gudu CG16719

Table IX. GSEA pathways leading edge gene subsets for GO Collection, BP Subcategory

pathway	size	leadingEdge
GO_NERVOUS_SYSTEM_PROCESS	26	aru Nlg2 dlg1 ck Nlg4 Adar kis Dcp-1 sgg mbt fz4 mgl sima CG9328 IRSp53 dsh Galphai Sh rdgA
GO_POSITIVE_REGULATION_OF_ RNA_BIOSYNTHETIC_PROCESS	59	myo Su(dx) ewg Lim3 skd Fhos mle Hnf4 CtBP Pdp1 kis cnc Gmap Dp Spt5 CASK pAbp pan CG4751 dom TfIIFalpha CG7154 CG6770 mtd Sirt1 fz4 sima pros Nufip dsh vg
GO_POSITIVE_REGULATION_OF_ TRANSCRIPTION_BY_RNA_ POLYMERASE_II	44	Su(dx) ewg Lim3 skd Fhos mle Hnf4 CtBP Pdp1 kis cnc Dp Spt5 CASK pan CG4751 TfIIFalpha mtd Sirt1 sima pros Nufip dsh vg CG43658 Ino80
GO_REGULATION_OF_TRANSPORT	47	Su(dx) SK dlg1 Nlg4 Rtnl1 milt mle CG5059 kis scrib cnc Dp CASK sgg Sik2 pan Fim chb Mdr50 Sirt1 sima dsh Galphai Sh rdgA Pp1-87B Ubqn Odc1
GO_BIOLOGICAL_ADHESION	37	CG32066 tty Nlg2 dlg1 Nlg4 Dcp-1 scrib CASK sgg chb mbt fz4 CG9328 IRSp53 Kank Pka-C3 LanA CG6066 mrj Pp1-87B noc sog CG31915 zip emc peb pico Rac2 N Tis11 tna Smox
GO_NEGATIVE_REGULATION_ OF_MOLECULAR_FUNCTION	49	myo Mkp3 Su(dx) dlg1 Adar skd CG32264 Dcp-1 scrib Dp sgg Sik2 pan TfIIFalpha CG2182 CG6770 Sirt1 pros dsh Galphai mrj rdgA Pp1-87B bip2 Ubqn Pdk Usp7
GO_NEGATIVE_REGULATION_OF_ CELLULAR_BIOSYNTHETIC_ PROCESS	65	myo Su(dx) ewg Lim3 skd Fhos mle Hnf4 CtBP Pdp1 kis cnc Gmap Dp Spt5 CASK pAbp pan CG4751 dom TfIIFalpha CG7154 CG6770 mtd Sirt1 fz4 sima pros Nufip dsh vg
GO_NEGATIVE_REGULATION_OF_ NUCLEOBASE_CONTAINING_ COMPOUND_METABOLIC_PROCESS	68	myo Su(dx) ewg Lim3 skd Fhos mle Hnf4 CtBP Pdp1 kis cnc Gmap Dp Spt5 CASK pAbp pan CG4751 dom TfIIFalpha capu CG7154 RpII215 CG6770 mtd Sirt1 fz4 sima pros Nufip dsh vg CG43658 Ino80 Usp7 salr Lim1 Not1 emc pum nonA peb Mrtf CG8519 Hr4 N Tis11 fs(1)h tna Smox sqd XRCC1
GO_REGULATION_OF_CELLULAR_ LOCALIZATION	31	dlg1 Nlg4 Rtnl1 milt mle CG5059 kis scrib Dp CASK sgg pan Fim sima dsh Galphai Sh rdgA Pp1-87B

GO_REGULATION_OF_CELL_ POPULATION_PROLIFERATION	63	myo Mgstl dlg1 Rtnl1 Adar Rox8 mle CtBP Dcp-1 scrib Dp CASK pan CG7154 CG2182 Jarid2 CG6770 Sirt1 mgl sima Kank pros dsh Galphai Sh LanA mrj Pp1-87B Mi-2 Odc1 noc sog sty CG8545 Lim1 disco-r emc peb Srpk79D l(1)G0320 RpL23 CkIIbeta Mrtf MYPT-75D Rac2 N Tis11 tna Smox
GO_PYRIMIDINE_CONTAINING_ COMPOUND_CATABOLIC_PROCESS	1	CG8349
GO_PYRIMIDINE_CONTAINING_ COMPOUND_BIOSYNTHETIC_ PROCESS	1	CG8349
GO_NUCLEOSIDE_SALVAGE	1	CG8349
GO_NUCLEOSIDE_CATABOLIC_ PROCESS	1	CG8349
GO_NUCLEOBASE_METABOLIC_ PROCESS	1	CG8349
GO_NUCLEOBASE_CONTAINING_ SMALL_MOLECULE_CATABOLIC_ PROCESS	1	CG8349
GO_NUCLEOBASE_CONTAINING_ SMALL_MOLECULE_ BIOSYNTHETIC_PROCESS	1	CG8349
GO_GLYCOSOL_COMPOUND_ CATABOLIC_PROCESS	1	CG8349
GO_GLYCOSOL_BIOSYNTHETIC_ CATABOLIC_PROCESS	1	CG8349
GO_CELLULAR_METABOLIC_ COMPOUND_SALVAGE	1	CG8349

4. Discussion

4.1 PCR Contributed to the Majority of Duplicated Reads

Duplicate reads can be caused by PCR duplicates that occur during the library preparation process, or by duplicates that occur during the sequencing process. Duplicates can confound the estimation of the transcript abundance if not properly handled and removed [28]. Fortunately, duplicate reads can be removed using bioinformatics tools such as MarkDuplicates in the Picard tool suite. The MarkDuplicates step of the pipeline was tested with two different option flags that could be set to remove all types of duplicates, such as PCR and sequencing duplicates, or a flag that could be set to only remove sequencing duplicates. Depending on the flag that was set, the step would result in different numbers of surviving reads following the removal of duplicates. Table IV shows the counts and percentages of the surviving reads for each type of flag. As observed in Table IV, the samples contained high counts of PCR duplicates as a significantly lower number of reads remained following the removal of all possible types of duplicates across all samples. In comparison to the raw input read quantities, roughly 10% of reads were viable for quantification across all samples.

The FastQC results were used as another form of validation for the presence of PCR duplicates in the reads after trimming. Metrics containing information about sample GC% content can be compared to the expected GC% content of *D. melanogaster* to demonstrate levels of contamination. In Figure 2, the per base sequence content graph output by FastQC for the BPA-exposed sample, A_2_S123_L004_R1_001, can be observed. The GC% content for this sample can be found by summing the average of the black and blue lines observed in Figure 2, and the combined GC% content appears to be around ~20-30%. In comparison, the expected

median GC% content in *D. melanogaster* is roughly 42% [29]. The expected GC% content of *D. melanogaster* is significantly higher than what is observed in this sample's reported GC% content, indicating the possibility of partial contamination, despite reads mapping to the indexed reference *D. melanogaster* genome in downstream analysis.

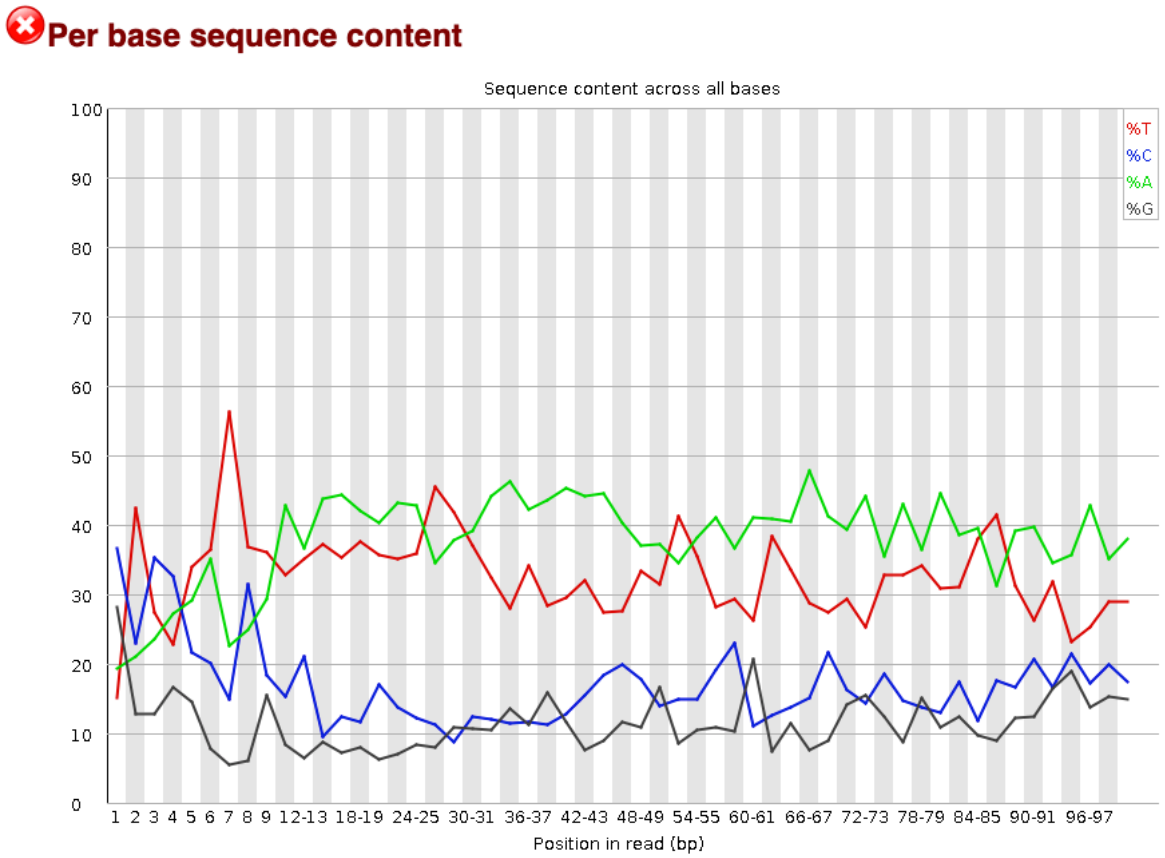

Figure 2. Per base sequence content for sample A_2_S123_L004_R1

Figure 3 contains a graph of the per sequence GC content output by FastQC for the same BPA-exposed sample, A_2_S123_L004_R1_001. The blue line is representative of the expected GC% content for *D. melanogaster*. It can be clearly seen that the GC count per read does not remain aligned with the expected theoretical distribution after the first 12 bases, and this could be due to contaminating sequences, such as different species, adapters, mitochondrial/rRNA, and/or overrepresented sequences. Multiple peaks in a GC content graph such as those observed in

Figure 3, are usually an indication of contamination.

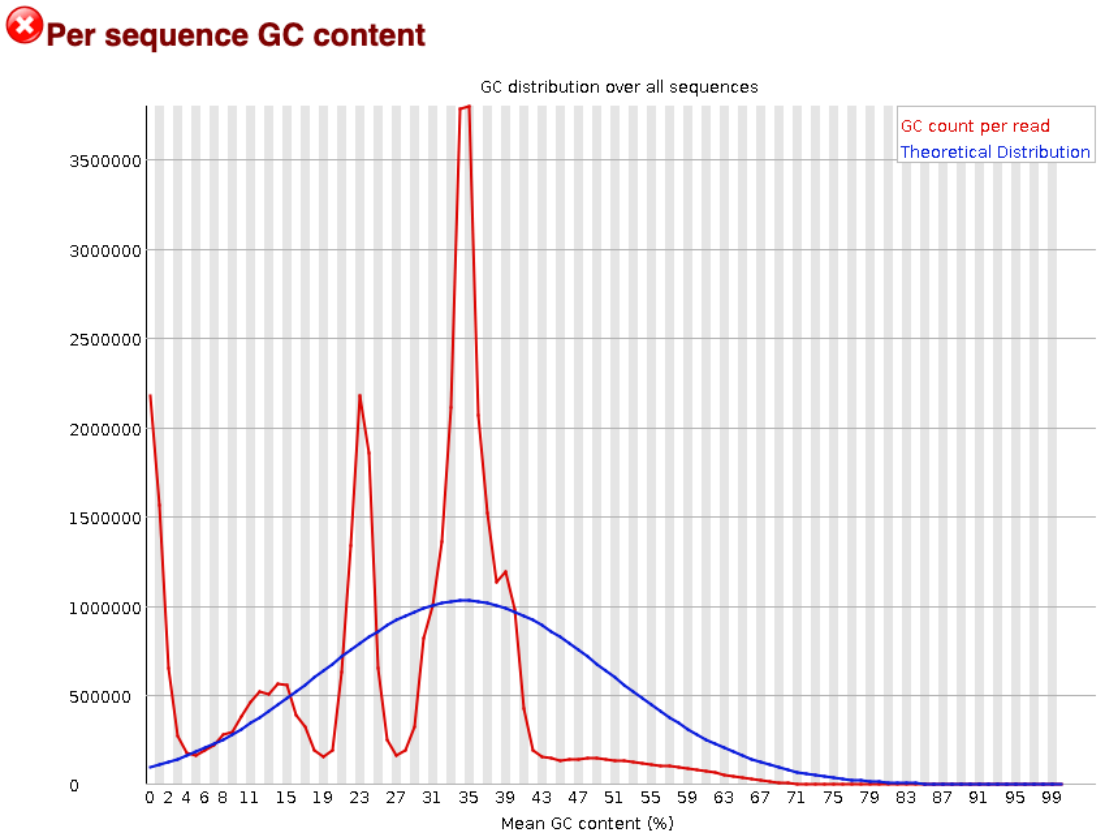

Figure 3. Per sequence GC content for sample A_2_S123_L004_R1

Further confirmation of the presence of duplicates was found by investigating sequence duplication levels. Figure 4 contains a graph of the sequence duplication levels output by FastQC for the same BPA-exposed sample A_2_S123_L004_R1_001. The blue line is representative of the total percent of sequences and reads in the sample. It can be seen that according to Figure 4, the percent of sequences remaining if the present reads were deduplicated would only be 5.32%. The high levels of sequence duplications could be due to a low complexity library used during preparation, too many cycles of PCR amplification during library preparation, or could also be due to too little of starting material and target messenger RNA (mRNA) material for sequencing.

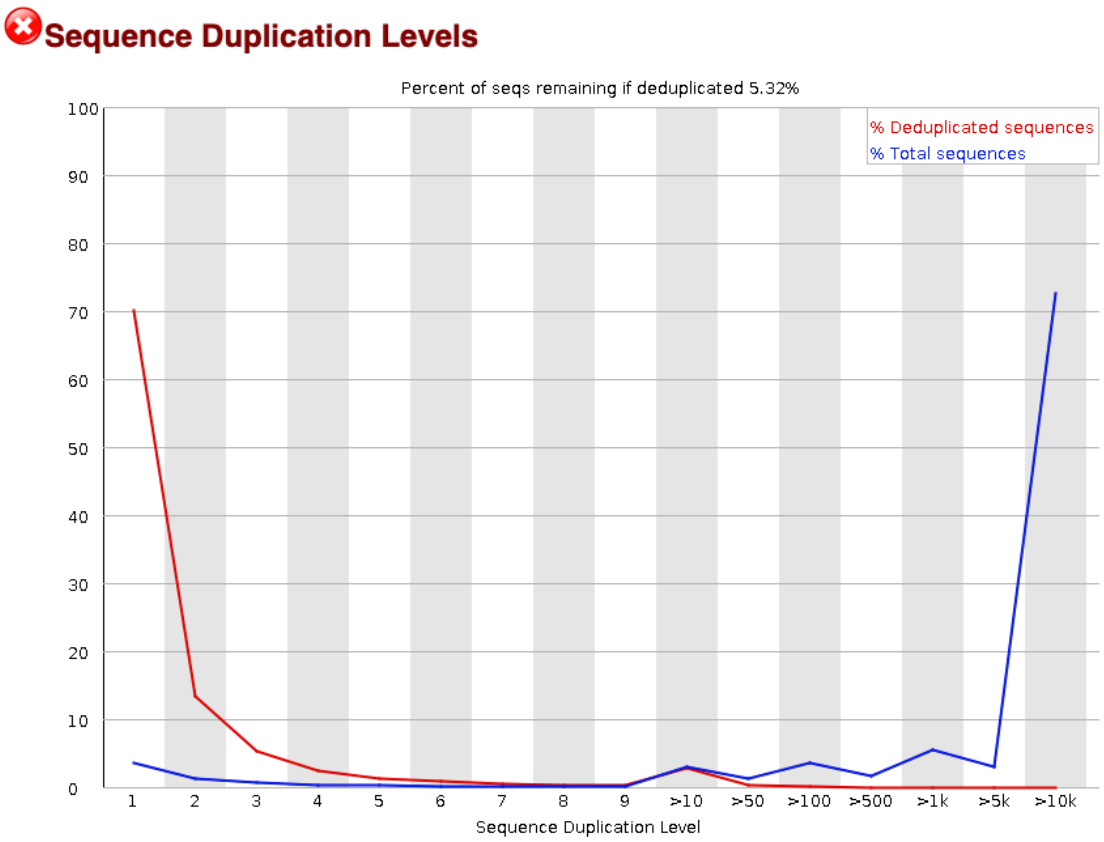

Figure 4. Sequence duplication levels for sample A_2_S123_L004_R1

An additional metric that was investigated output by the FastQC results for the same BPA-exposed sample A_2_S123_L004_R1_001 was overrepresented sequences. Two of the top 3 overrepresented sequences, "GTCCTTTCGTACTAAAATATCACAATTTTTTAAAGATAGAAACCAACCTG" and "GTCCAACCATTCATTCCAGCCTTCAATTAAAAGACTAATGATTATGCTAC", accounted for 6.74% and 2.67% of overrepresented sequences in the sample, respectively. A BLASTn search was performed on these two sequences [30]. Figure 5 shows the top BLAST hit for the first overrepresented sequence, which had 100% identity and 100% coverage matching the *D. melanogaster* mitochondrial complete genome. Figure 6 shows a BLAST hit that was among the top hits for the second overrepresented sequence, in which the query overrepresented sequence

had 100% identity and 100% coverage matching a different *Drosophila* species' ribosomal RNA. The contaminating sequences, while *Drosophila*, are not the target mRNA material for analysis. It would be ideal to avoid the amplification of this non-target material through the use of a ribo-depletion kit that can be used during the library preparation step. The use of a ribo-depletion kit would also minimize the waste of sequencing resouces.

Figure 5. Top BLAST hit for first overrepresented sequence in sample A_2_S123_L004_R1

Figure 6. BLAST hit for second overrepresented sequence in sample A_2_S123_L004_R1

While DE analysis could have continued with only the removal of sequencing duplicates and the pipeline would still have been able to correctly identify affected pathways, it was decided to remove all duplicates, due to the high levels of duplicated sequences and to generate stronger results. It appears that there were a substantial number of PCR duplicates in all samples; thus, it was essential that this PCR bias was accounted for with the conservative removal.

4.2 BPA Exposure Caused Up-regulated Gene Expression Involved in Axon Guidance and Neuron Development

DESeq2 performed differential expression analysis by performing analysis on the count files generated from the transcript quantification step. Table V summarizes the significant results exported by DESeq2. The results of Table V indicate that the majority of the differentially expressed genes that had p-adjusted values less than 0.01 were genes that were up-regulated by BPA exposure. Furthermore, there are no outliers in this subset of statistically significant results,

indicating that the data is not skewed to show more up or down-regulated genes. As noted in Table V, of these significant differentially expressed genes, there were 0 with low counts, providing more evidence that the differences in expression detected by DESeq2 were entirely due to genes that had high differences in counts.

Figure 7 contains a heatmap of the normalized count matrix from DESeq2 data and displays levels of expression among differing sample conditions for the top 50 differentially expressed genes. In addition, Figure 7 includes a size factor and condition labeling scheme in its legend.

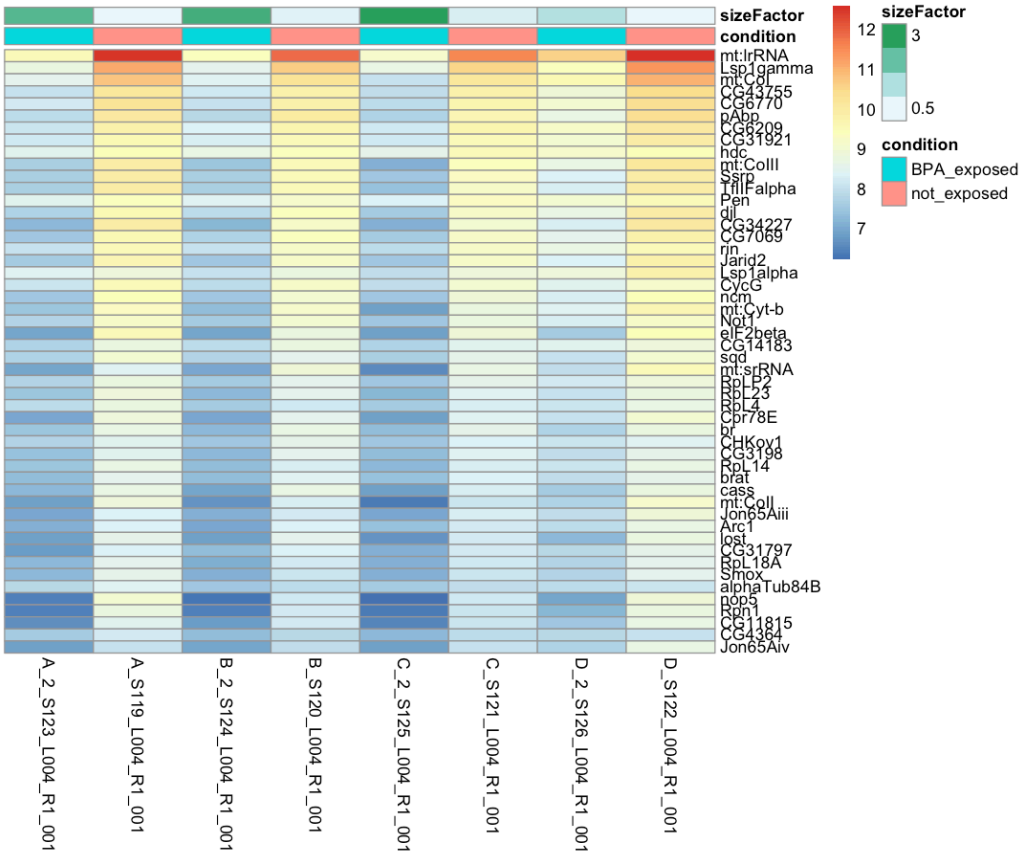

Figure 7. Heatmap of Normalized Count Matrix

The statistically significant genes (genes with p-adj < 0.01) were output into a `.csv` file and ordered by ascending p-adj values. The subset of differentially expressed genes included

1040 genes that were exported into the `.csv` file based upon the p-adj threshold of < 0.01. The `.csv` file can be found in the appendix. One of the notable top 50 genes differentially expressed based upon normalized counts, *Smox*—this gene is involved with axon guidance, mushroom body development, and neuron development [31]–[33]. *Smox* being present in this heatmap of the normalized count matrices is aligned with cellular and behavioral phenotypes observed in the lab of Dr. Kimberly Mulligan.

Log2 fold change was plotted in a histogram in Figure 8 based upon the 1040 statistically-significant differentially expressed genes output from DESeq2. 50 bins were created based upon the log2 fold change value on the x-axis, and frequency was plotted on the y-axis for each bin. It can be observed that most genes exported in the significant results file experienced a positive fold change since the histogram plot has heavier weight to the right-side of the plot; however, higher frequency of the fold-changes of these genes were on the scale of 1. Figure 8 demonstrates that very few genes experienced extremely high (greater than 4) or extremely low (less than -4) fold change. However, of the 10 genes that experienced extremely high log2 fold change, 4 of the genes have known neurodevelopmental significance in *D. melanogaster*. The 4 genes, *aru, myo, tou,* and *Exn,* are listed in Table X. Since the histogram from Figure 8 was created with input data from all 1040 statistically significant differentially expressed genes, the fact that the neurodevelopmentally-related genes are among the genes with extremely high log2 fold change, indicates that these genes involved in neurodevelopment stand out as being differentially regulated by BPA, even when examining all differentially expressed genes regardless of biological context.

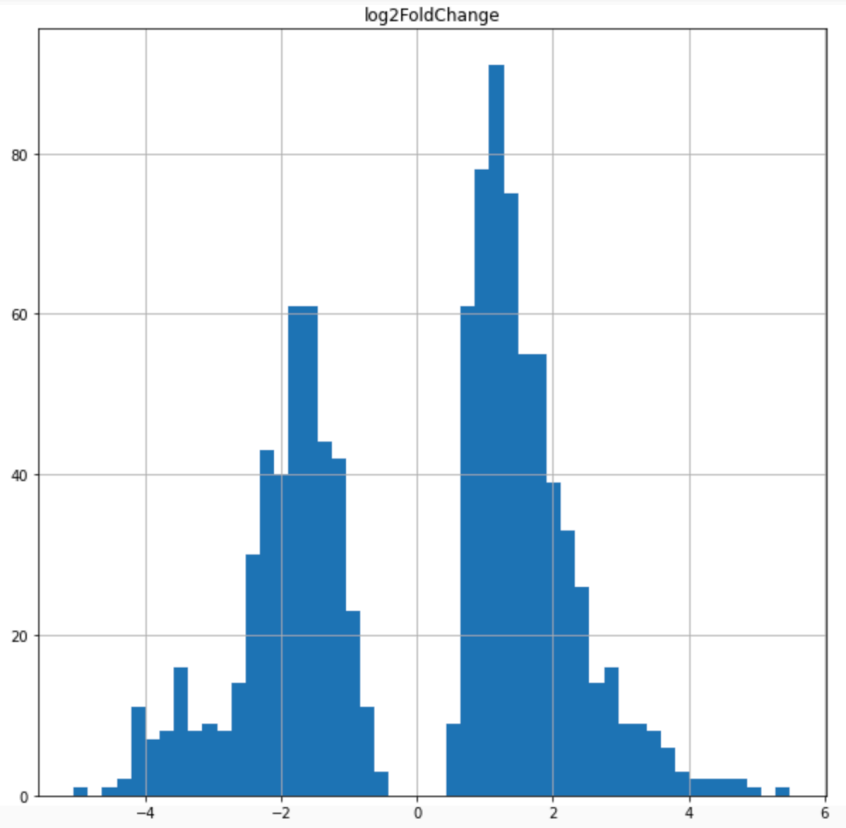

Figure 8. Histogram plot of log2 fold change and frequency for 1040 genes (padj < 0.01)

Table X. Top 10 differentially expressed genes with log2 fold changes greater than 4

symbol	log2 fold change
rdgA	5.494587
Osi15	4.930355
myo	4.792463
tou	4.746440
aru	4.569626
CG32066	4.526561
Mkp3	4.417412
Su(dx)	4.246549
Exn	4.152533
lncRNA:CR45919	4.102425

4.3 Molecular Pathways Involved with Axon Guidance and Neuron Development Affected by BPA Exposure

Gene set enrichment analysis provides insight into the specific molecular pathways and genes subject to change following exposure to an experimental condition. A GSEA table was generated by fgsea and is shown in Figure 9. This figure displays gene ranks, NES, and statistical significance for the top 10 up-regulated and top 10 down-regulated pathways for GO gene set collection, CC subcategory from MSigDB for *D. melanogaster*. The GSEA table provided insight as to which molecular processes were influenced by BPA exposure, and multiple pathways showed direct association with neurologically significant pathways. All of the pathways that were listed in the table had significantly low p-values, and almost half of the top up-regulated pathways were of neurological significance. Some of the pathways listed that were of note were GO_SYNAPSE, GO_PRESYNAPSE, GO_POSTSYNAPSE, GO_GLUTAMATERGIC_SYNAPSE, and GO_DENDRITE_TERMINUS as these pathways have direct known association with neurological processes as reported in the Molecular Signature database [20]–[23]. The presence of these pathways—which are involved with synapses, neurotransmitters and the endocrine system—further supports the observed cellular and behavioral impacts of BPA as an EDC observed in the lab of Dr. Mulligan. Based upon the results observed in Figure 9, it can be concluded that BPA affects neurological molecular pathways in developing *D. melanogaster* larvae, in a manner that is of statistical significance with significance being based upon LFC ranking.

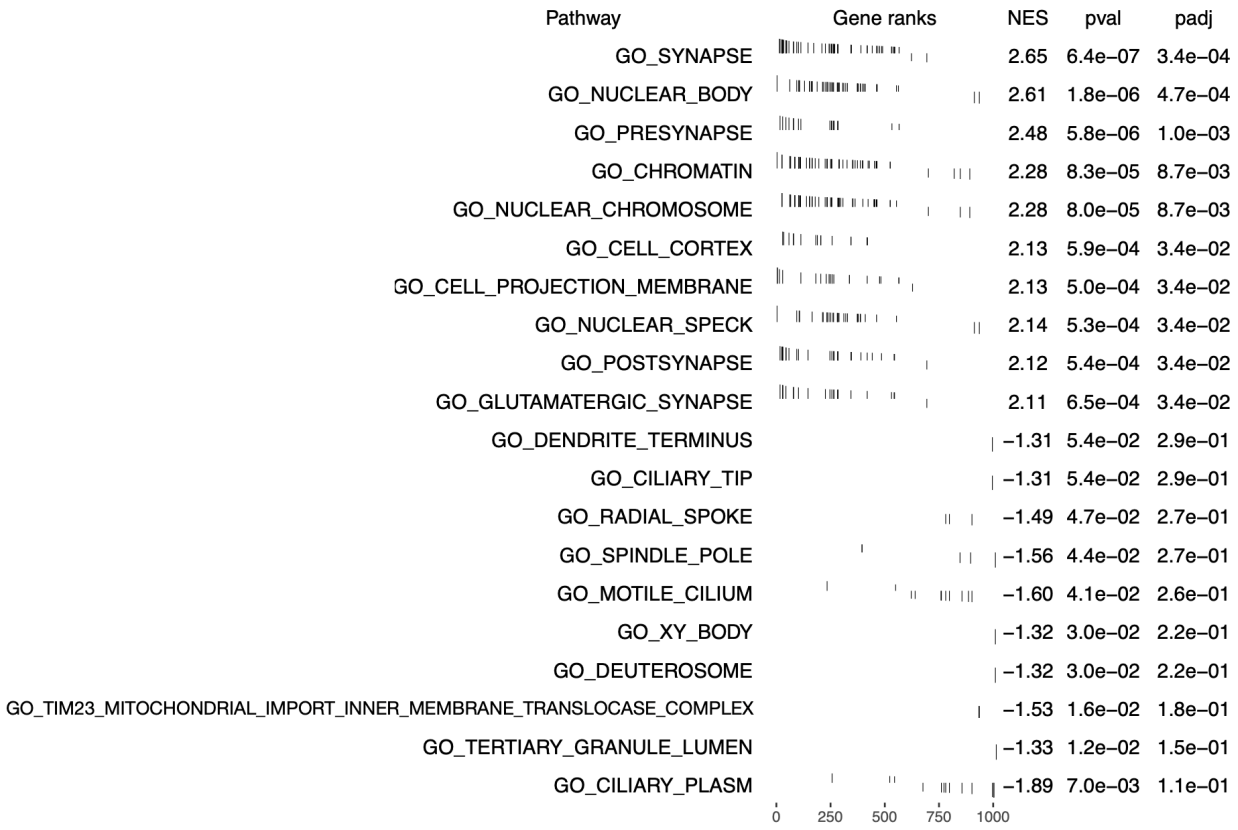

Figure 9. GSEA Table for the GO collection, CC subcategory of gene sets in D. melanogaster

Enrichment plots of the notable molecular pathways related to neurodevelopmental pathways are included below. The enrichment plots help denote the ranking of their leading edge genes as those genes included in the subset are represented by the black dashes before the maximum enrichment score, denoted by the dashed red line across the plot in Figures 10-14**.** Given the size of the leading edge subsets for each pathway listed in Table VIII, one can get an idea of how many genes are included in these pathways based upon the number of black dashes on the "rank" line of Figures 10-14. Note in Figure 14, the number of black dashes, or the number of genes in the pathway are notably lower in comparison to what is observed in Figures 10-13. This is also reflected in Table VIII in the "size" column.

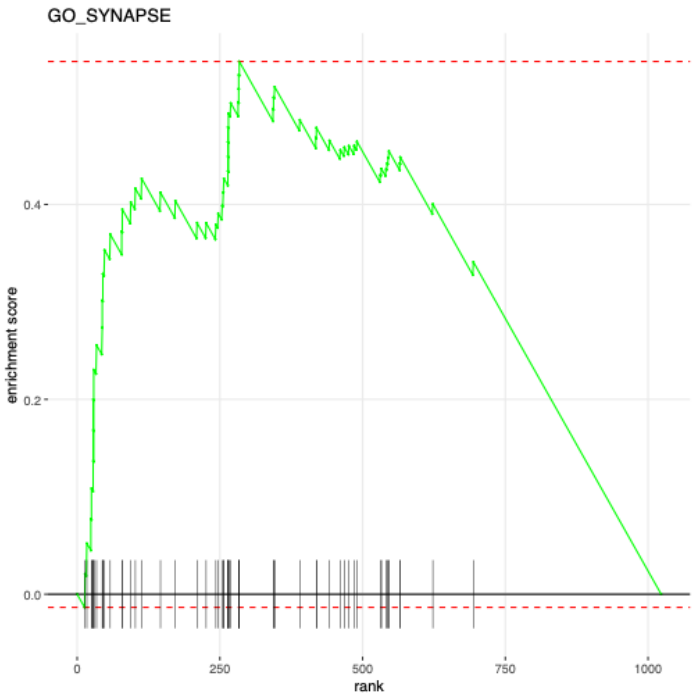

Figure 10. Enrichment plot for the GO_SYNAPSE pathway

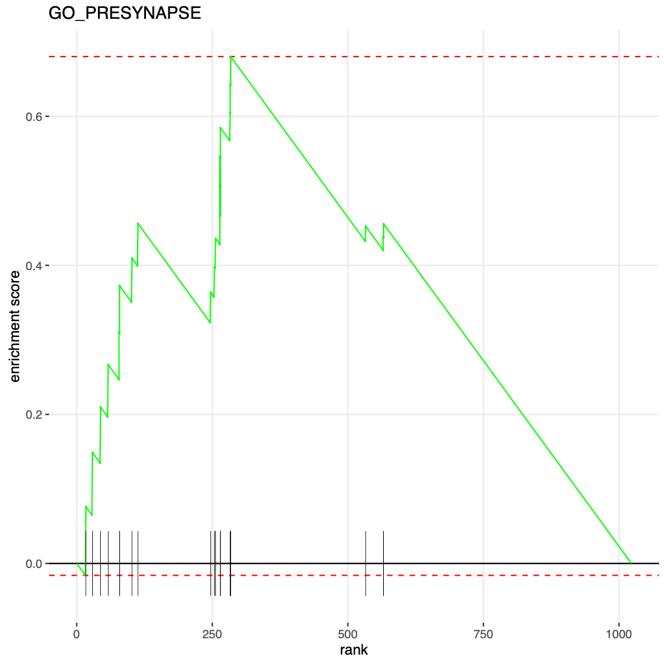

Figure 11. Enrichment plot for the GO_PRESYNAPSE pathway

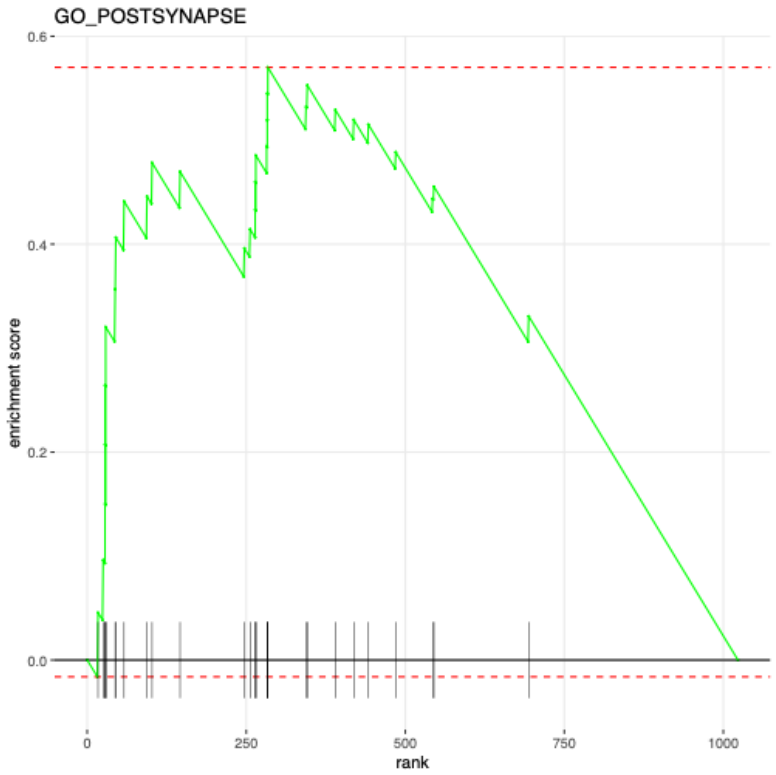

Figure 12. Enrichment plot for the GO_POSTSYNAPSE pathway

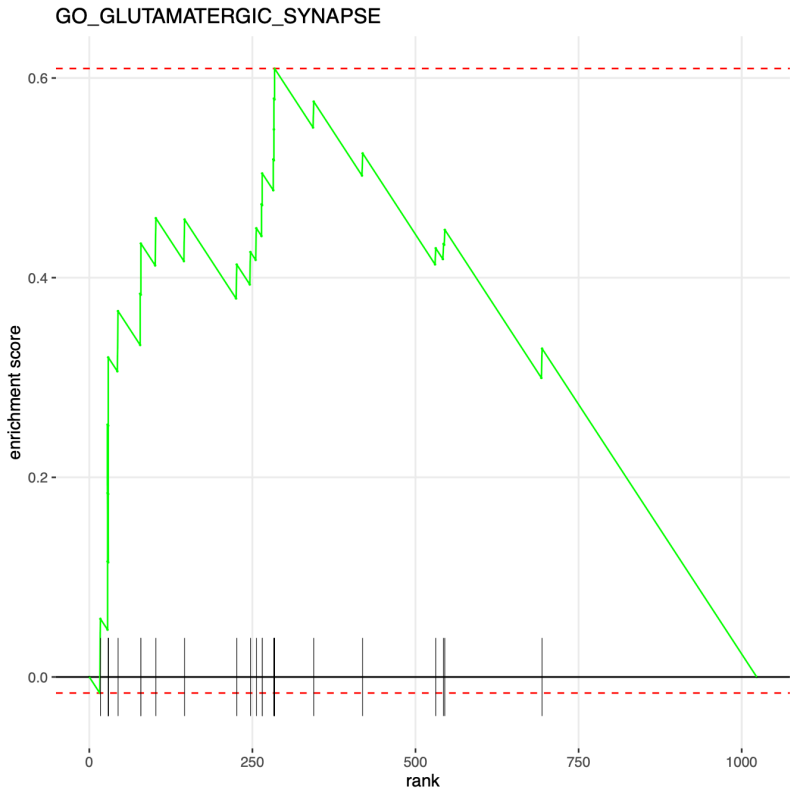

Figure 13. Enrichment plot for the GO_GLUTAMATERGIC_SYNAPSE pathway

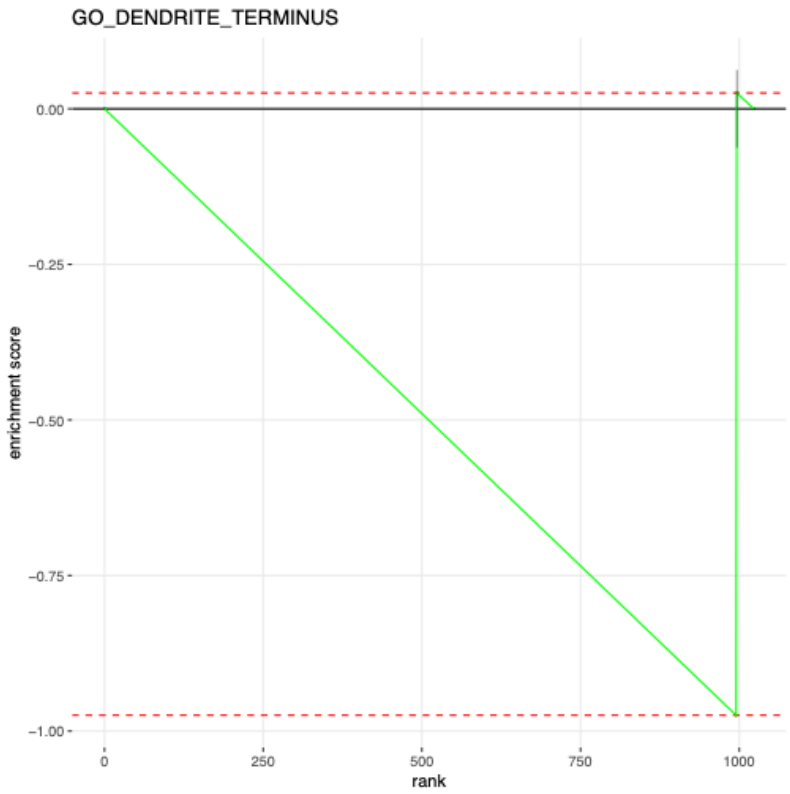

Figure 14. Enrichment plot for the GO_DENDRITE_TERMINUS pathway

To further investigate using GSEA, the gene set collection was refined to only include those pathways GO gene set, and was further refined to only include the BP subcategory. When refining to the GO BP subset, the following GSEA table is generated in Figure 15.

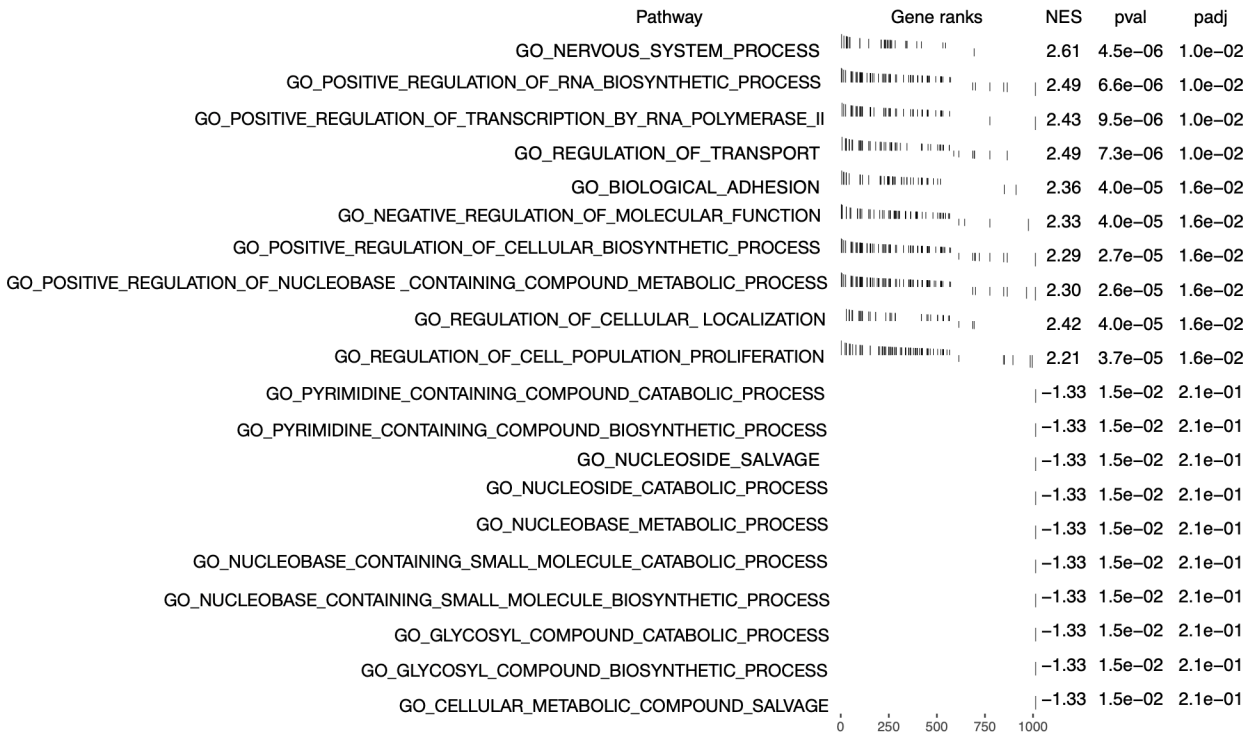

Figure 15. GSEA Table for the GO collection, BP subcategory of gene sets in D. melanogaster

The GSEA table listed in Figure 15 shows that the most up-regulated pathway when focusing on this particular gene set collection is the GO_NERVOUS_SYSTEM_PROCESS pathway, an additional neurodevelopmentally-relevant pathway identified by the analysis. In Figure 16 below, the enrichment plot for this pathway can be found.

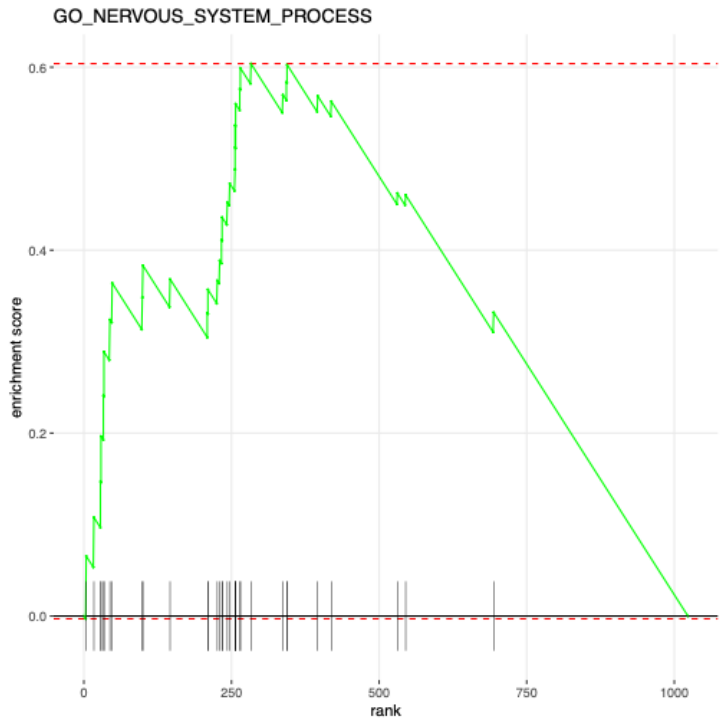

Figure 16. Enrichment plot for the GO_NERVOUS_SYSTEM_PROCESS pathway

The data in Figure 15 clearly indicates that further refining the collection of genes to the GO BP subset and performing GSEA with the genes observed in the significant results revealed an additional pathway with neurodevelopmental significance. Narrowing the collection of gene sets to the GO Collection instead of looking over all possible processes affected helped bring new insights to the neurodevelopmentally-related pathways affected by BPA-exposure. The GSEA step was able to reveal neurodevelopmentally-relevant pathways, with direct association with synapses, dendrites, and endocrine signaling.

4.4 Leading Edge Genes Include Neurodevelopmentally Significant Genes

The leading edge genes are the genes in a pathway that contributed the most to the enrichment score. Genes that are present in multiple leading edge subsets for different pathways

are likely susceptible to differential expression following exposure to an experimental condition. Leading edge genes were organized into a table based upon frequency in the pathways listed in the GSEA tables from Figures 9 and 15, from the GO CC and BP gene sets. The counts of the genes reflect the number of statistically significant molecular pathways the genes are involved in; for genes with higher frequency, one can infer that these genes are subject to differential expression following exposure to BPA. As observed in Table XI, eight of the genes that appeared in multiple leading gene subsets were previously identified to be involved in functions that affect cellular and behavioral outcomes, including axon guidance, courtship, and locomotion (*dlg1, CASK, Adar, nonA, mbt, Smox, myo, Rtnl1*) [34]–[41].

Two genes of particular interest are among the highest frequency— *dlg1* and *CASK*. The gene *dlg1* is important for nervous system development and courtship behavior [34], [42]. In addition, *CASK* encodes a protein important for memory, synaptic transmission at the neuromuscular junction, and courtship behavior [43], [35], [44]. These two genes have established neurodevelopmental pathway significance and are among the most frequent of the leading edge genes; their frequency indicates that neurodevelopmental pathways are subject to change following BPA exposure. Another notable gene that is among the most frequent, is *dsh*. The gene *dsh* is the dishevelled gene that was originally identified in mutant flies [45]. In a study conducted by Srahna et. al 2006, *dsh* was shown to be needed for dorsal cluster neuron axon extension and stabilization in wild-type *Drosophila* [46]. The identification of leading edge genes assist in identifying the molecular underpinnings of the cellular and behavioral phenotypes identified in the Mulligan research group [6].

All of the leading edge genes listed in Table XI provide explicit support and molecular evidence that BPA exposure influences neurodevelopmentally-related genes and pathways; 39 genes of known neurologically-significant function are differentially expressed. The GSEA step of the RNA-Seq analysis pipeline was able to identify specific neurodevelopmentally-related pathways and genes that has not been done in other studies involving RNA-Seq analysis in *D. melanogaster* following BPA exposure.

Table XI. Neurologically significant leading edge genes that occurred in more than one pathway

Gene	Frequency	Gene	Frequency
dsh	16	Smox (mushroom body development)	5
CASK (male courtship behavior)	12	Rtnl1 (axon/axonal cone growth)	5
dlg1 (male courtship behavior)	12	Adar (male courtship behavior)	4
Pp1-87B	12	chb	4
mle	10	Dcp-1	4
scrib	9	Lim1	4
Sh	9	nonA (male courtship behavior)	4
cnc	8	Srpk79D	4
Galphai	8	LanA	3
Nlg4	8	mbt (mushroom body development)	3
pros	8	N	3
sgg	8	SK	3
dom	7	tou	3
kis	7	aru	2
Lim3	6	Atx2	2
milt	6	jing	2
Nlg2	6	noc	2
myo (mushroom body development)	5	Rac2	2
pAbp	5	zip	2
peb	5		

4.5 Future Directions

Future projects can incorporate the usage of the RNA-Seq pipeline deployed here to determine differential expression in *D. melanogaster* resulting from environmental exposures and/or genetic risk factors associated with NDDs. Given that NDDs have both environmental and genetic etiologies that may work in concert with one another, exploring the influence of BPA in different genetic backgrounds may provide further insight into how BPA influences the pathophysiology of NDDs. In addition, this pipeline may also be used for execution on *D. melanogaster* samples following exposure to other environmental chemicals. There are thousands of environmental chemicals in use that have undergone little to no toxicological testing. This pipeline, combined with *D. melanogaster* as a model, provides a rapid risk assessment strategy to identify chemicals that may confer risk of NDDs. Of immediate relevance to the current project, this pipeline could be used to compare the impact of BPA-analogs commonly used in BPA-free products—in order to compare their impact to that of BPA. Such a study would help determine if BPA-free products are indeed a safer choice than BPA-containing products.

Possible areas of further research can include further gene set enrichment analysis that can hopefully help identify additional molecular pathways that are neurodevelopmentally-relevant. Additional research may include a deeper investigation into the differential expression of the leading edge genes listed in Table XI. A future project may include a specific focus on these genes in wild-type larvae since multiple molecular pathways affected by BPA include these genes. In addition, it is of interest to determine the relative influence of BPA-analogs on

molecular pathways, with a targeted interest in investigating their effects on neurological pathways.

REFERENCES

[1] S. E. Arambula, S. M. Belcher, A. Planchart, S. D. Turner, and H. B. Patisaul, "Impact of Low Dose Oral Exposure to Bisphenol A (BPA) on the Neonatal Rat Hypothalamic and Hippocampal Transcriptome: A CLARITY-BPA Consortium Study," *Endocrinology*, vol. 157, no. 10, pp. 3856–3872, Oct. 2016.

[2] A. D. Henriksen, A. Andrade, E. P. Harris, E. F. Rissman, and J. T. Wolstenholme, "Bisphenol A Exposure in utero Disrupts Hypothalamic Gene Expression Particularly Genes Suspected in Autism Spectrum Disorders and Neuron and Hormone Signaling," *Int. J. Mol. Sci.*, vol. 21, no. 9, Apr. 2020.

[3] G. Schönfelder, W. Wittfoht, H. Hopp, C. E. Talsness, M. Paul, and I. Chahoud, "Parent bisphenol A accumulation in the human maternal-fetal-placental unit.," *Environ. Health Perspect.*, vol. 110, no. 11, pp. A703–A707, Nov. 2002.

[4] S. K. Tiwari, S. Agarwal, L. K. S. Chauhan, V. N. Mishra, and R. K. Chaturvedi, "Bisphenol-A impairs myelination potential during development in the hippocampus of the rat brain," *Mol. Neurobiol.*, vol. 51, no. 3, pp. 1395–1416, 2015.

[5] S. K. Tiwari et al., "Inhibitory Effects of Bisphenol-A on Neural Stem Cells Proliferation and Differentiation in the Rat Brain Are Dependent on Wnt/β-Catenin Pathway," *Mol. Neurobiol.*, vol. 52, no. 3, pp. 1735–1757, Dec. 2015.

[6] U. Nguyen et al., "Exposure to bisphenol A differentially impacts neurodevelopment and behavior in Drosophila melanogaster from distinct genetic backgrounds," *Neurotoxicology*, vol. 82, pp. 146–157, Jan. 2021.

[7] T. P. Stein, M. D. Schluter, R. A. Steer, L. Guo, and X. Ming, "Bisphenol A Exposure in Children With Autism Spectrum Disorders," *Autism Res.*, vol. 8, no. 3, pp. 272–283, 2015.

[8] J. Baio, "Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014," *MMWR Surveill. Summ.*, vol. 67, 2018.

[9] M. R. Herbert, "Contributions of the environment and environmentally vulnerable physiology to autism spectrum disorders," *Curr. Opin. Neurol.*, vol. 23, no. 2, pp. 103–110, Apr. 2010.

[10] M. D. Rand, "Drosophotoxicology: the growing potential for Drosophila in neurotoxicology," *Neurotoxicol. Teratol.*, vol. 32, no. 1, pp. 74–83, Feb. 2010.

[11] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, p. 550, 2014.

[12] "Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data." https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[13] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinforma. Oxf. Engl.*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014.

[14] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype," *Nat. Biotechnol.*, vol. 37, no. 8, Art. no. 8, Aug. 2019.

[15] M. Pertea, D. Kim, G. Pertea, J. T. Leek, and S. L. Salzberg, "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie, and Ballgown," *Nat. Protoc.*, vol. 11, no. 9, pp. 1650–1667, Sep. 2016.
[16] H. Li et al., "The Sequence Alignment/Map format and SAMtools," *Bioinforma. Oxf. Engl.*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.
[17] "Picard Tools - By Broad Institute." https://broadinstitute.github.io/picard/.
[18] S. Anders, P. T. Pyl, and W. Huber, "HTSeq--a Python framework to work with high-throughput sequencing data," *Bioinforma. Oxf. Engl.*, vol. 31, no. 2, pp. 166–169, Jan. 2015.
[19] "Fast gene set enrichment analysis | bioRxiv." https://www.biorxiv.org/content/10.1101/060012v3.full.
[20] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, Jun. 2011.
[21] A. Subramanian et al., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci.*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005.
[22] "GSEA." https://www.gsea-msigdb.org/gsea/index.jsp (accessed Apr. 21, 2021).
[23] igor, *igordot/msigdbr*. 2021.
[24] J. Köster and S. Rahmann, "Snakemake—a scalable bioinformatics workflow engine," *Bioinformatics*, vol. 34, no. 20, pp. 3600–3600, Oct. 2018.
[25] A. Larkin et al., "FlyBase: updates to the Drosophila melanogaster knowledge base," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D899–D907, Jan. 2021.
[26] A. D. Yates et al., "Ensembl 2020," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D682–D688, Jan. 2020.
[27] M. D. Adams et al., "The Genome Sequence of Drosophila melanogaster," *Science*, vol. 287, no. 5461, pp. 2185–2195, Mar. 2000.
[28] Z. Fang and X. Cui, "Design and validation issues in RNA-seq experiments," *Brief. Bioinform.,* vol. 12, no. 3, pp. 280–287, May 2011.
[29] "Drosophila melanogaster (ID 47) - Genome - NCBI." https://www.ncbi.nlm.nih.gov/genome?term=vih&cmd=DetailsSearch (accessed May 03, 2021).
[30] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J Mol Biol*, vol. 215, no. 3, pp. 403–410, Oct. 1990.
[31] L. Parker, J. E. Ellis, M. Q. Nguyen, and K. Arora, "The divergent TGF-beta ligand Dawdle utilizes an activin pathway to influence axon guidance in Drosophila," *Dev. Camb. Engl.*, vol. 133, no. 24, pp. 4981–4991, Dec. 2006.
[32] X. Zheng et al., "TGF-beta signaling activates steroid hormone receptor expression during neuronal remodeling in the Drosophila brain," *Cell*, vol. 112, no. 3, pp. 303–315, Feb. 2003.
[33] X. Zheng, C. T. Zugates, Z. Lu, L. Shi, J. Bai, and T. Lee, "Baboon/dSmad2 TGF-β signaling is required during late larval stage for development of adult-specific neurons,"

EMBO J., vol. 25, no. 3, pp. 615–627, Feb. 2006.

[34] C. Mendoza-Topaz et al., "DLGS97/SAP97 Is Developmentally Upregulated and Is Required for Complex Adult Behaviors and Synapse Morphology and Function," *J. Neurosci.*, vol. 28, no. 1, pp. 304–314, Jan. 2008.

[35] J. B. Slawson, E. A. Kuklin, A. Ejima, K. Mukherjee, L. Ostrovsky, and L. C. Griffith, "Central regulation of locomotor behavior of Drosophila melanogaster depends on a CASK isoform containing CaMK-like and L27 domains," *Genetics*, vol. 187, no. 1, pp. 171–184, Jan. 2011.

[36] Y. A. Savva et al., "Auto-regulatory RNA editing fine-tunes mRNA re-coding and complex behaviour in Drosophila," *Nat. Commun.*, vol. 3, no. 1, Art. no. 1, Apr. 2012.

[37] M. B. Sokolowski, "Drosophila: genetics meets behaviour," *Nat. Rev. Genet.*, vol. 2, no. 11, pp. 879–890, Nov. 2001.

[38] J. Melzig *et al.*, "A protein related to p21-activated kinase (PAK) that is involved in neurogenesis in the Drosophila adult central nervous system," *Current Biology*, vol. 8, no. 22, pp. 1223–1226, Nov. 1998.

[39] L. Parker, J. E. Ellis, M. Q. Nguyen, and K. Arora, "The divergent TGF-β ligand Dawdle utilizes an activin pathway to influence axon guidance in Drosophila," *Development*, vol. 133, no. 24, pp. 4981–4991, Dec. 2006.

[40] T. Awasaki, Y. Huang, M. B. O'Connor, and T. Lee, "Glia instruct developmental neuronal remodeling through TGF-β signaling,", *Nat. Neurosci.*, vol. 14, no. 7, pp. 821-823, Jun. 2011.

[41] K. Rao et al., "Spastin, atlastin, and ER relocalization are involved in axon but not dendrite regeneration," *Mol Biol Cell*, vol. 27, no. 21, pp. 3245–3256, Nov. 2016.

[42] T. Ohshiro, T. Yagami, C. Zhang, and F. Matsuzaki, "Role of cortical tumour-suppressor proteins in asymmetric division of Drosophila neuroblast," *Nature*, vol. 408, no. 6812, pp. 593–596, Nov. 2000.

[43] B. R. Malik, J. M. Gillespie, and J. J. L. Hodge, "CASK and CaMKII function in the mushroom body α'/β' neurons during Drosophila memory formation," *Front. Neural Circuits*, vol. 7, p. 52, 2013.

[44] M. Sun et al., "Genetic interaction between Neurexin and CAKI/CMG is important for synaptic function in Drosophila neuromuscular junction," *Neurosci. Res.*, vol. 64, no. 4, pp. 362–371, Aug. 2009.

[45] M. Sharma, I. Castro-Piedras, G.E. Simmons, and K. Pruitt, "Dishevelled: A masterful conductor of complex Wnt signals," *Cell. Signal.*, vol. 47, pp. 52-64, Jul. 2018.

[46] M. Srahna, M. Leyssen, C. M. Choi, L. G. Fradkin, J. N. Noordermeer, and B. A. Hassan, "A signaling network for patterning of neuronal connectivity in the Drosophila brain," *PLoS Biol.*, vol.4, no. 11, p. e348, Oct. 2006.

APPENDIX

Trimmed FastQC Files

https://www.dropbox.com/s/sh0q4hvnk31hq25/A_2_S123_L004_R1_001_fastqc.html?dl=0

https://www.dropbox.com/s/m02v0z4s7vswtqm/A_S119_L004_R1_001_fastqc.html?dl=0

https://www.dropbox.com/s/uh1dtm4zgdyw1z6/B_2_S124_L004_R1_001_fastqc.html?dl=0

https://www.dropbox.com/s/3cu92tfcs9a0888/B_S120_L004_R1_001_fastqc.html?dl=0

https://www.dropbox.com/s/vkzpk37ipj6irn3/C_2_S125_L004_R1_001_fastqc.html?dl=0

https://www.dropbox.com/s/rfz16wxp5pnq8ju/C_S121_L004_R1_001_fastqc.html?dl=0

https://www.dropbox.com/s/74ph1ljqrp4ys4h/D_2_S126_L004_R1_001_fastqc.html?dl=0

https://www.dropbox.com/s/br2kr5tohna0sky/D_S122_L004_R1_001_fastqc.html?dl=0

DESeq2 CSV File

https://www.dropbox.com/s/yrza7ndljr7mw2p/significant_results.csv?dl=0