Transcriptome Analysis of Neuronal and Muscle Tissue in Smn Mutant Drosophila melanogaster

Transcriptome Analysis of Neuronal and Muscle Tissue in *Smn* Mutant *Drosophila melanogaster*

Hansine Heggeness

A Thesis in the Field of Bioinformatics

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

May 2021

Abstract

Spinal muscular atrophy (SMA) is a devastating human neurodegenerative disease caused by a defect in the gene *SMN1* resulting in a loss of the protein Survival Motor Neuron (SMN). The model organism *Drosophila melanogaster* has a conserved homologue of the human *SMN1* gene, *Smn*, and its knockdown produces SMA-like phenotypes. SMN's role in the cell is not fully understood; despite being ubiquitous and cell lethal it appears to target motor neurons in the SMA disease phenotype, suggesting tissue specific function. This study investigates the tissue-specificity of *Smn* function and its role in *Drosophila* muscle and neuronal tissue. We used RNA-Seq analysis of muscle and CNS to investigate transcriptome changes when *Smn* is knocked down in a model that mimics the human SMA phenotype. We found many differentially expressed genes and changes in differential exon usage. We explored the enriched GO terms and KEGG pathways associated with the gene sets from each respective tissue. To further investigate *Smn*'s role in the muscle we demonstrated a dose-dependent model of SMA and found a set of genes whose expression profiles are correlated with the amount of *Smn* present in the tissue. We compared our results from each tissue to what is currently known about SMN and to our lab's previous findings. Our results show that many aspects of *Smn*'s role in the cell are tissue-specific; each tissue shows distinct differentially expressed genes and differentially used exons. Not only are the genes themselves tissue-specific but the significantly enriched biological processes and pathways associated with them were as well.

Dedication

This thesis is dedicated to my fiancé, Sam, whose unwavering love, support, and encouragement was with me in every step of my studies. Thank you for believing in me and making sure I believe in myself.

I would also like to dedicate this work to my family for their support and love.

Acknowledgments

I would like to thank my supervisor Dr. Van Vactor for his guidance, instruction, and advice that was instrumental throughout the entire project. I would also like to thank Jae-Yoon Jung for answering all of my coding questions.

Table of Contents

List of Tables

List of Figures

Chapter I

Introduction

SMA Overview

Children with spinal muscular atrophy (SMA) have trouble supporting their heads and, if the disease is severe, they die within days or weeks of being born (GARD, 2018). In fact, SMA is the leading genetic cause of death of infants (FDA, 2019). The patient's age at the onset of the disease corresponds to the severity of the disease. Different severities of symptoms are classified into "types" 0-4, but all have the same underlying genetic cause. Type 0 leads to death in the first days or weeks of the infant's life, while type 4 is adult onset SMA where muscle weakness occurs but life expectancy is not affected (GARD, 2018). General progressive muscle weakness can manifest as difficulty with breathing, swallowing, or standing (GARD, 2018). Unlike some genetic conditions, such as sickle cell or Tay Sachs, SMA is not more likely to affect a particular ethnic group. In one study done in the United States, the carrier frequency for the disease was 1 in 54 of a pan-ethnic population (Sugarman et al., 2012). SMA has a reported incidence of affecting between 1 in 10,000 and 1 in 11,000 live births. In 95% of cases of SMA, the patient has a genetic deletion in the gene survival of motor neuron 1 (*SMN1*) such that the gene is unable to code for the protein SMN (Kolb & Kissel, 2011). Humans have two copies of the survival of motor neuron gene, *SMN1* and *SMN2*, though *SMN2* is usually largely nonfunctional. This nonfunction is owing to a C-to-T substitution at the 5′ end of exon 7 found in *SMN2* that results in the exclusion of exon 7. However, a small quantity

of proteins are still able to be properly transcribed which can amount to about 10% of the normal amount of SMN (Kolb & Kissel, 2011). The severity of SMA is determined by the amount of SMN protein available (Lefebvre et al., 1997) lower levels result in more severe phenotypes. Most people have multiple copies of *SMN2* and this copy number affects the amount of SMN that is available to the cells (Kolb & Kissel, 2011). The majority of *SMN1* mutations are deletions, usually in exon 7, preventing the gene from producing SMN. These mutations account for the autosomal recessive inheritance of the disease, both parents must have a faulty copy and pass these faulty copies to their offspring to result in a child with SMA. The other 5% of cases occur due to point mutations that change *SMN1* enough such that it cannot produce SMN. In these cases, the patient usually has inherited one faulty deletion copy of *SMN1* and had this point mutation in the other copy. SMA affects the motor neurons in the anterior horn of the spinal cord where the low level of SMN causes the neurons to gradually die off. When the muscles no longer receive neural impulses from the brain they atrophy; this causes the muscle weakness symptom in SMA. The severity of the disease is determined by the amount of SMN protein present and thus determined by how much *SMN2* can make up for the loss of function of *SMN1* (Kolb & Kissel, 2011). Patients with a higher copy number of *SMN2* generally have fewer symptoms because the multiple copies are able to generate more SMN. Other factors that may contribute to the overall severity of the disease are a topic of current research.

There are several therapies in development for treatment of SMA, many of which focus on increasing SMN expression. Several antisense oligonucleotides (ASO) have been made to try to affect the *SMN2* exon 7 splicing. The FDA in the United States and

regulatory bodies in a few other countries have approved one ASO, Spinraza (Nusinersen), as a treatment for all levels of SMA (Groen et al., 2018). Spinraza targets an intronic element upstream of exon 7, leading to increased inclusion of exon 7 and increased levels of full-length SMN protein (Shorrock et al., 2018). It is still being tested for efficacy in patients of various disease stages and is not a full cure for the disease. Spinraza was tested in pre-symptomatic infants with *SMN1* deletion. While there was some success in infants achieving many age-appropriate developmental milestones, observations one year into treatment report only 30% of infants standing unaided. This result indicated that while early intervention is helpful it is not a cure (Shorrock et al., 2018). One drawback of ASO usage is that these treatments are administered by intrathecal injection into the cerebral spinal fluid at several points for direct delivery to the central nervous system (CNS) and the injections need to be repeated over time (Groen et al., 2018). These injections themselves can be painful and time-consuming with side effects from administration like lower back pain and post-lumbar puncture syndrome as well as side effects due to the drug itself. These drug administration complications and the fact that despite increasing SMN levels Spinraza does not fully cure the disease mean researchers are looking for more options for therapies. Other scientists are exploring small molecule therapies which also affect exon 7 splicing (Groen et al., 2018) and could be administered orally, like the recently FDA-approved Risdiplam (Singh et al., 2020). Risdiplam promotes the inclusion of exon 7 by modulating its splicing. There is no current consensus on the mechanism by which it works to such a specific degree and there are a few known off-target effects. The advantage of small molecule therapies is the ease of access and treatment: it is much easier to take an oral medication than to have an

intrathecal injection. Also, this drug can be shipped and stored at ambient temperature, decreasing the cost. Risdiplam has demonstrated the potential use for other small molecule therapies for SMA and their combined use with other non-SMN-targeting treatments. One further SMN targeted option is gene therapy; the results so far have been promising but more expanded research will be needed (Groen et al., 2018). Other therapies do not target SMN: Olesoxime is an oral drug that is found to have some neuroprotective properties when used on type 2 and type 3 patients (Groen et al., 2018). Some therapies target other cellular aspects of SMA and possible interactors with SMN. Supplementary drugs target muscle symptoms in patients and aim to improve muscle function (Groen et al., 2018). The SMN independent therapies could be used in combination with ASOs or other SMN dependent therapies to address many aspects of disease recovery and provide customizable treatment plans for individuals with varying severity and in different disease stages. By further understanding SMN's role in the cell we understand more of its interactors; finding targets downstream in pathways affected by SMN loss provides options for therapies to bring these pathways back to their normal levels, possibly circumventing a need for SMN in the cell.

Current Understanding of SMN

Despite the focus on motor neurons in the SMA disease model, SMN is a ubiquitous protein found in all human tissues (Groen et al., 2018). SMN is localized in the cytosol and the nucleus with a presence in the axon (Pagliardini et al., 2000). SMN is most well known and most studied for its involvement in the assembly of the spliceosome and in the synthesis and assembly of ribonucleoproteins, specifically small nuclear

ribonucleoproteins (snRNPs) (Matera & Wang, 2014). Five different snRNPs make up the spliceosome, an important cellular machine responsible for the removal of introns from pre-mRNA in a process called splicing (Matera & Wang, 2014). Splicing removes the noncoding portions of RNA to form the mature mRNA that is processed during the translation of proteins. Therefore, snRNPs are important to create functional proteins. The SMN protein complex, which includes SMN and several gemins, regulates the cytoplasmic maturation phase of snRNP assembly (Matera & Wang, 2014). The SMN dissociates from snRNP when it is imported into the nucleus. SMN appears to be particularly important for the formation of snRNPs that remove U12 introns; some proteins containing U12 introns are found incorrectly spliced in *Drosophila Smn* mutant models (Lotti et al., 2012). SMA mice with low levels of SMN have been shown to have increases in aberrant splicing events with RNA-Seq analysis in various tissues, supporting a view that improper splicing events are a mechanism of the disease (Doktor et al., 2017).

SMN also plays a role in many other cell functions, such as mRNA translation and trafficking as well as cytoskeleton assembly (Chaytow et al., 2018). SMN has been found localized in axons, dendrites, and the growth cone, and has been linked to proper transport and then local translation of mRNA. It is also found to be in association with ribosomal units for local translation and regulation of the mTOR pathway (Chaytow et al., 2018). These functions may be a reason why motor neurons can be so drastically affected by low levels of SMN despite its expression in all cells. SMN is also important in endocytotic pathways, particularly in the neuromuscular junction (NMJ) (Dimitriadi et al., 2016). Disruption of activity and physical defects of the NMJ are an important

phenotype of SMA. SMN is also thought to have a role in autophagy (Chaytow et al., 2018) and it has some function in ubiquitin pathways and in mitochondria (Chaytow et al., 2018). In many of these processes the exact extent to which SMN is involved is still unknown. Additionally, the exact reason for motor neurons' acute susceptibility to low levels of SMN, as opposed to other cell types that are not affected as severely, is also unknown. *Smn* RNA-Seq-focused work in *Drosophila* by Garcia et al. and others have used whole larval tissues. These important findings can be further refined by targeting a specific tissue with additional observation of the SMA specific outcome of *Smn* loss.

Drosophila melanogaster as a Model Organism

The *SMN* gene family is highly conserved, which makes its study using model organisms possible. *Drosophila melanogaster* are an extremely useful model organism for researching human disease. *Drosophila* have a short life cycle and are easy to breed and maintain, an advantage over larger organisms. Many fly lines with various mutations exist; these fly lines can be ordered by mail and maintained for relatively low cost. *Drosophila* are a great model organism to use in genetics because the genetics of the species makes it possible to create lines with multiple mutations over balancer chromosomes, thus enabling researchers to maintain lines over multiple generations (Tolwinski, 2017). While flies have a smaller number of chromosomes (4 pairs) than humans (23 pairs), flies have approximately 15,500 genes whereas humans have about 22,000 genes. Additionally, about 60% of the fly genome is homologous to the human genome (NIH, 2012). When using *Drosophila* there are multiple genetic tools available to researchers that make this model organism especially attractive to work with. The Gal4

UAS system is a genetic tool used in fruit fly research that enables researchers to selectively express genes in specific tissues of the organism (Duffy, 2002). This enables the overexpression or loss of function of a natural gene, or even the expression of a gene not found in wild type *Drosophila*, like channelrhodopsin or GFP.

Drosophila SMA Models

The *Drosophila* gene *Smn* is a conserved homologue of human *SMN1*, however, unlike the human *SMN1* and *SMN2*, the *Drosophila* genome contains only a single copy of the gene. While the full loss or deletion of *Smn* is as lethal in flies as it is in humans and other animals, the partial loss of function of the *Smn* gene results in a phenotype that mimics the human phenotype of the disease including such symptoms as viability, locomotion, and NMJ defects (Chang et al., 2008). Using the Gal4 UAS system an *Smn* loss of function RNA interference (RNAi) can be expressed specifically in various tissues, for example, neuronal tissue or muscle, while leaving the other cell types unchanged. Many *Smn* RNAi lines have been made which can be used to look at various levels of disease severity phenotypes ranging from mild to acute (Spring et al., 2019). Many other model organisms are unsuccessful in mimicking the Type 3 and 4 phenotypes of SMA; mouse models, for instance, are often either severely affected or unaffected, leaving out important intermediate forms of SMA (Spring et al., 2019). Additionally, the different life stages of *Drosophila*, larval and adult, can act as different models that can be used to cover a range of SMA types (Spring et al., 2019). The tissue-specific RNAi knockdown of *Smn* in the CNS or muscle of *Drosophila* is able to recapitulate the phenotype of the disease without compromising viability of the

larva. *Smn* tissue-specific knockdown in both muscle and neuronal tissue causes pupal lethality, with stronger *Smn* knockdowns resulting in a higher percentage of mortality in the population (Chang et al., 2008). Smn has been found to congregate in the postsynaptic side of the NMJ colocalizing with Discs Large, a common postsynaptic marker, and in muscle fibers. *Smn* expression in both muscle and CNS is required for normal NMJ activity (Chang et al., 2008). Targeting specific modifiers of *Smn* in different tissues rescues different dimensions of the disease phenotype. For example, activation of FGF signaling in muscles rescues synaptic defects caused by *Smn* RNAi (Sen et al., 2011). It is clear that both muscle and neuronal tissue expression are important in *Smn* function. Using *Drosophila* as a model organism avoids ethical and technical issues involved in obtaining and working with human samples and larger model organisms. *Drosophila* provide a simpler conserved system to study with readily available tools primed for investigations of tissue-specific activity.

Objectives of This Study

It is known that SMN is required in all cell types (Groen et al., 2018). What is less understood is the degree to which SMN has different function in different cell types. Given that the SMA disease phenotype is largely based on muscle weakness relating to the NMJ and loss of function in motor neurons, we understand that the impact of SMN's loss is not the same across all cell types and seems to impact motor neurons more severely. This study is therefore focused on loss of *Smn* specifically in the neuronal and muscle tissue to observe the differences in these tissue-specific transcriptomes with respect to differential gene expression as well as differential exon usage. RNA-Seq analysis found genes that are up- or down-regulated in the disease model and identified

processes and pathways that are affected in the specific tissues. We observed confirmation of previous interactors found in genetic studies done in *Drosophila* and significant changes in the aforementioned *Smn* related pathways. The phenotype of SMA is contingent on the amount of SMN available to the cell and that most investigations have looked at neuronal effects of SMN loss. Therefore, we are interested in which genes and their processes and pathways are affected in a similar dose-dependent manner to *Smn,* specifically in muscle, using a set of *Drosophila* lines that present a series of SMN expression. This unique observation produces a gradient model of SMA analogous to the human dose-dependent phenotype. A better understanding of *Smn* and its function in each tissue provides a first steppingstone that in the future could facilitate work around possible druggable pathways and targets for muscle tissue to be used in combination with neuronal therapies to support and improve SMA care.

Chapter II

Materials and Methods

Fly Strains

The fly strains used were *w*; *P{UAS-WIXZ}, w*; *P{UAS-SmnRNAi-C24}*, *w*; *Df(3L)SmnX7,P{UAS-SmnRNAi-C24}/TM6B*, and the driver lines *Mhc-Gal4/TM6B* and *w*; *P{w*$^{+mC}$*=GAL4-elav.L}3*. Driver line females of *Mhc-Gal4/TM6B* for muscle assay or *w*; *P{w*$^{+mC}$*=GAL4-elav.L}3* for CNS assay were crossed to males of *w*; *P{UAS-WIXZ}, w*; *P{UAS-SmnRNAi-C24}*, or *w*; *Df(3L)SmnX7,P{UAS-SmnRNAi-C24}/TM6B* to obtain the intended genotypes for dissection seen in Table 1. Control lines were the empty construct made in the same genetic background as our transgenic lines.

Table 1. Fly Mating Scheme

Muscle	Paternal strain	Maternal strain	Genotype under analysis
PWIZ	*P{UAS-WIXZ}*	*Mhc-Gal4*	*Mhc-Gal4 / PWIZ*
C24	*P{UAS-Smn$^{RNAi-C24}$}*	*Mhc-Gal4*	*Mhc-Gal4 /P{UAS-Smn$^{RNAi-C24}$}*
X7C24	*Df(3L)Smn^{X7},P{UAS-Smn$^{RNAi-C24}$}/TM6B*	*Mhc-Gal4*	*Mhc-Gal4 / Df(3L)Smn^{X7},P{UAS-Smn$^{RNAi-C24}$}*
CNS	**Paternal strain**	**Maternal strain**	**Genotype under analysis**
PWIZ	*P{UAS-WIXZ}*	*P{Gal4-elav.L}3*	*P{Gal4-elav.L}3 / PWIZ*
X7C24	*Df(3L)Smn^{X7},P{UAS-Smn$^{RNAi-C24}$}/TM6B*	*P{Gal4-elav.L}3*	*P{Gal4-elav.L}3 / Df(3L)Smn^{X7},P{UAS-Smn$^{RNAi-C24}$}*

RNA Preparation

RNA preparation was done under direction of Takakazu Yokokura in the Van Vactor lab at the Okinawa Institute of Science and Technology (OIST). Late third instar larva of the intended genotypes were dissected in ice cold PBS. For the muscle sample,

each biological group consisted of 5 larval pelts with internal organs, tracheal tissues, and axon bundles and fibers removed. For the central nervous system sample each biological group consisted of 200 CNS and VNCs pulled from the larva and pooled into each individual sample. Total RNA was extracted and the quality was tested.

Pooled samples were homogenized in TriPure Isoration Reagent using a Polytron homogenizer following the manufacture's instruction protocol for RNA extraction procedure with minor modification. Crude RNA was treated with rDNase set (Macherey-Nagel, Duren, Germany) and subsequently purified using the NucleoSpinRNA Clean-up XS kit (Macherey-Nagel, Duren, Germany). Purified total CNS RNA was quantified using spectrophotometry (NanoDrop; ThermoFisher Scientific, Waltham MA, USA); quality of the total RNA was tested by using microfluidic analyzer, Agilent RNA 6000 Nano kit (Agilent Technologies, Waldbronn, Germany), and by examining expression levels of *Usp7*, *GAL4*, *elav*, *repo*, *SK*, *RpL32*, *Pen*, *Smn*, *tkv*, and *wg* genes by qRT-PCR. Purified total muscle RNA was quantified using Qubit Fluorometer ThermoFisher Scientific, Waltham MA, USA); quality of the total RNA was tested by using microfluidic analyzer, Agilent 4200 TapeStation (Agilent Technologies, Santa Clara, CA, USA. Samples were sent to the OIST sequence center.

Sequence Library Construction and Next Generation Sequence - CNS

Three micrograms of total RNA were used as template for construction of one biological replicate of the sequence library. After purification, mRNA was incubated at 95^{o} C for 3 minutes to fragment the RNA. cDNA was synthesized using fragmented RNA as a template, repaired, adenylated at its 3' end, and ligated with adaptors. cDNA copy number contained in the libraries was measured using the BioRad QX200 Droplet

Digital PCR system (ddPCR). To prepare the sequence library, the cDNA library was amplified by running 10 cycles of PCR. After amplification, we quantified copy number in each sequence library by using ddPCR Library Quantification for Illumina TruSeq (BioRad, Hercules, CA) with QX200 Droplet Digital PCR system (BioRad, Hercules, CA). Libraries were run on Illumina HiSeq 2000 for sequence analysis. We used TruSeq RNA sample Prep Kit v2 – Set A for sample preparation, and TruSeq PE Cluster Kit v3 – cBot – HS, HiSeq cBot Manifold, TruSeq HiSeq Accessory v3, TruSeq Multiplex Sequence Primer, TruSeq SBS Kit v3 – HS (200 cycles) and HiSeq PE Flow cell v3 for sequencing. Three libraries were run on one lane of flow cell.

Sequence Library Construction and Next Generation Sequence - Muscle

One microgram of total RNA was used as template for constructing one biological replicate of the sequence library. After isolation of mRNA using NEBNext Poly(A) mRNA Magnetic Isolation Module (New England BioLabs Inc, Ipswich MA), sequence libraries were constructed using NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England BioLabs Inc). We chose libraries enriched with 10 PCR cycles to avoid over amplification. Libraries were run on Illumina NovaSeq 6000 system for sequence analysis. Libraries were loaded on reagent cartridges with S2 flow cell (NovaSeq 6000 S2 Reagent kit V1 (300 cycles)) and pair-ended reads were run on Illumina NovaSeq 6000 for sequencing.

RNA-Seq Pipeline

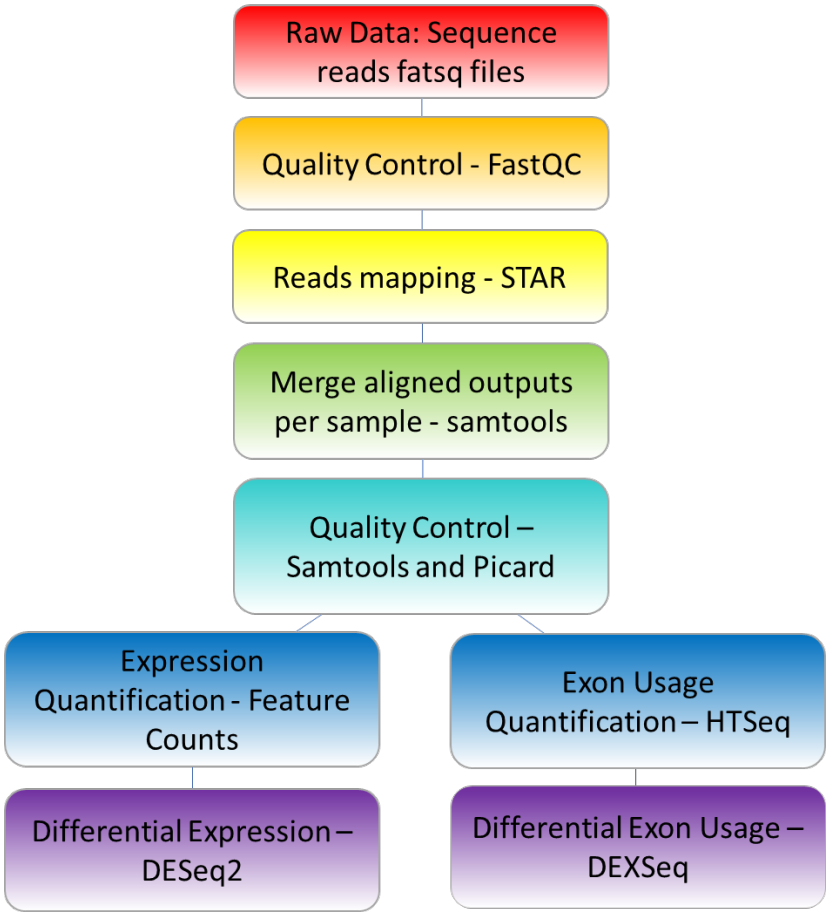


Figure 1. RNA-Seq Workflow

Summary of the workflow taken to process the raw fasta files through to the point of Differential Expression analysis or Differential Exon Usage analysis.

Raw paired-end fastq files obtained from the sequencing facility were uploaded to O2 High Performance Computer Cluster and put into the pipeline shown in Figure 1. The O2 cluster is used because it is able to process the files and programs faster than a single computer and it has all the tools needed readily available to load into the work

environment. Multiple jobs can be run at a time, further cutting down the amount of time it takes to run through the data. Files are unzipped and then reads are processed for quality control by FastQC v0.11.3 (Andrews, 2010). FASTQC takes in raw fastq files and outputs a graphical report with various metrics graded. These metrics include basic statistics, per base sequence quality, sequence quality scores, per base sequence content, per sequence "gc" content, per base n content, sequence duplication levels, overrepresented sequences, adapter content, and kmer content. All files were found to be within acceptable and expected ranges. Then paired files that passed quality control were aligned against a reference genome using Spliced Transcripts Alignment to a Reference v2.7.3a (STAR). The reference genome used for alignment was r6.33 (FB2020_02) from Flybase.net. STAR is a very fast mapper due to its two-step algorithm process of 1) seed searching and 2) clustering stitching and scoring; it "exhibits better alignment precision and sensitivity than other RNA-Seq aligners" (Dobin et al., 2013). Next the produced BAM files were merged per sample to make 1 file per biological replicate using samtools v1.9. These files are then indexed and processed for quality control by samtools and then by Picard v2.8.0. Finally the indexed BAM files are processed by featureCounts (Liao et al., 2014), a read summarization program from the subread/2.0.0 package. featureCounts takes as input the BAM files and a gene annotation file (gtf format) for mapping the reads. The gtf used was r6.33. This step outputs a file with read counts from the given samples for each gene. featureCounts counts reads at the gene level that uniquely map to a single location. The total read count for a gene is the sum of the reads of each exon that belong to that gene.

Once we have the counts file, we are able to download it to a local computer and upload the data into R and the IDE RStudio. R is a computer language and environment for statistical computation. It is available for free and widely used. The Bioconductor program DESeq2 (Love et al., 2014) was used to determine differential gene expression (DE). DESeq2 is a common and popular program for DE analysis. It takes as input a count matrix where the rows are genes and their counts and the columns are individual samples. Sample level quality control was performed using this package. The first step of quality control is count normalization which is necessary to perform gene count comparisons between samples. DESeq2 performs this using the median of ratios method. Next in the DESeq2 QC workflow comes sample-level QC using Principal Component Analysis (PCA) and hierarchical clustering methods. This enables the visualization of how well each of the biological replicates cluster together. Replicates of the same genotype should be similar to each other and cluster by sample group. Additionally, DESeq2 automatically performs gene level QC by filtering out genes with 0 counts in all samples, genes with single sample extreme count outliers, and genes with low mean normalized counts. Next, DESeq2 is run to analyze the differential expression of genes. Contrasts are created to outline which sample groups are to be compared and which group is the control vs the experimental group. Using this contrast, the program performs a Wald test to obtain the resulting p value, adjusted p value, and fold change of the genes' expression between the two contrasted sample groups. We set our p value cutoff to less than 0.05 for statistically significant differentially expressed genes. These steps were carried out on both the muscle dataset and the CNS dataset independently to obtain their individual significantly differentially expressed gene

lists. Once we have these significant gene lists, we are able to compare them and use them for further investigation.

We also looked at the differential exon usage (DEU) of our datasets using DEXSeq (Anders et al., 2012). DEXSeq is another R package; it is able to test for changes in the relative exon usage based on the experimental conditions. First, we convert the gtf reference file, the same reference used in the DE steps, to a .gff file. This step and the counting step use Python scripts in the DEXSeq package that utilize HTSeq (Anders et al., 2015), a Python package used to process high throughput sequencing data. We run the reference preparation and then run the counting step on the BAM files created earlier in the pipeline. We create the same contrast groups as in the DE analysis, comparing the *Smn* RNAi to control. DEXSeq performs normalization and dispersion estimation using the same methods as DESeq2. It then tests for differential exon usage by fitting a generalized linear model for each gene. We set our false discovery rate threshold to less than 0.05. We generated a list of significantly differentially used exons and the transcripts associated with these exons. From this list of transcripts, we found the associated genes and used this to create our list of DEU genes.

Dose Dependency Determination

In our muscle dataset we compared two different models of SMA severity using the C24 and X7C24 genotypes to test for dose dependency on *Smn* of the differentially expressed genes. We obtained the normalized count per sample and then from that the sample group average of normalized counts. We looked for the normalized count average pattern of Pwiz < C24 < X7C24 or X7C24 < C24 < Pwiz. If a gene fit this pattern it was considered to be dose-dependent.

In the DEU analysis we were also able to look at dose dependency. The DEXSeq program generates exon usage coefficients that are variance-stabilized transformed for each of the three samples. We used these coefficient values to compare the exon usage between samples and find the patterns that indicated dose dependance. From the list of significant exons that fit the pattern, we looked at the transcripts associated with these up or downregulated exons and matched them to their associated genes. We call this list our DEU dose-dependent genes.

Gene Set Overlap

Various gene lists were compared for overlap and significance. These lists were compared using the GeneOverlap R package which automates the process of performing Fisher's exact test on two gene lists and their overlap given a background genome (Shen, 2016). The lists compared were our RNA-Seq generated DE and DEU significant genes in muscle and in CNS. We also looked at stage specific gene lists L1, L2, and L3 compiled from Flybase and a list of 372 Smn genetic modifiers in *Drosophila* compiled from previous genome-wide functional screens in *Drosophila* by Sen (Sen et al., 2013) and Chang (Chang et al., 2008).

DIOPT Analysis

The identification of orthologs between model organism systems and humans is an important step in demonstrating models for human disease. To generate a list of significant human conserved genes we used the DRSC integrative ortholog prediction tool version 8 (DIOPT). This tool was created by the *Drosophila* RNAi Screening Center (DRSC) by Hu *et al* (Y. Hu et al., 2011) and can be found online at

http://www.flyrnai.org/diopt. They have created a user-friendly tool that compares the gene or genes of interest from any of a number of organisms (human, mouse, fly, worm, zebrafish, and yeast) to a variety of 16 ortholog databases (Ensembl Compara, HomoloGene, Inparanoid, Isobase, OMA, orthoMCL, Phylome, RoundUp, and TreeFam and more). The tool then gives a score and rank based on the number of databases that support a given gene pair relationship. Variations in the source ortholog tool algorithms and annotation databases account for the differences in the various ortholog predictions.

We input our lists of significant DE and DEU genes in muscle and in CNS in Flybase ID format. DIOPT's database is not as up to date as the Flybase database that we used to generate our DE and DEU gene lists and thus not all our genes were found in the DIOPT database. 14 genes were missing from the DE muscle set, 10 from the DE CNS set, 10 from the DEU muscle list, and 12 from the DEU CNS list. As these genes were not in the database, we could not check them for possible human orthologs. On the rest of the lists, we used the preset filter of removing orthologs with score less than 1 unless it was the only ortholog found for a specific gene.

GO and KEGG Enrichment

We further investigated our gene lists by looking at their known biological functions and pathways based on their associated gene ontology (GO) terms (Ashburner et al., 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa & Goto, 2000). We looked at both the global gene set enrichment and at specific significant list enrichment of these terms. To do this we used the clusterProfiler package in R (Yu et al., 2012). clusterProfiler takes in a gene list of interest and a background gene list performs statistical enrichment analysis using hypergeometric

testing. This can be done for GO terms (biological processes, cell component, and molecular function) as well as KEGG pathways. The associated GO terms come from the R package org.Dm.eg.db (Carlson, 2019) which contains genome wide annotations for *Drosophila melanogaster* based on Entrez ID identifiers. When possible, the GO term lists were simplified or reduced to remove redundant and similar GO terms in order to make the results more human readable and interpretable. The metric used is how "semantically similar" two GO terms are, using a cutoff of 0.6. These lists were then sent through the CateGOrizer version 3.218 online tool to further simplify and batch together terms into their ancestor GO classifications (Z.-L. Hu et al., 2008). To do this we used the Go_Slim2 library of 124 ancestor terms and the tool mapped our GO BP terms to this set of classifications. A simplified explanation is that each GO term is mapped back in its GO hierarchy until the program finds an ancestor listed in the GO_Slim2 list. These ancestor classifications are added up and we see the number of original terms that fall in that classification. It used the single counting method to limit repeats in the counting step. If a term maps to the same ancestor in more than one way, it only counts it as one hit. Once all terms were charted we graphed the percentage of terms related to each classification. Occasionally a GO term would not map to any ancestor term category in our selected GO_Slim2 list. These were not included in the classification analysis. Additional KEGG searches were performed on the KEGG website using the KEGG Mapper tool to search for gene hits in known pathways.

Chapter III

Results

Differentially Expressed Genes

Both muscle and CNS datasets were found to pass quality control with 300+ million reads and an average of 93% aligned in BAM in muscle and 230 million reads and an average of 95% aligned in BAM in CNS. The files were used for transcriptome analysis. Sample quality control using PCA and hierarchical clustering showed correlation between biological replicates. In the muscle dataset we looked at the transcriptome of three different genotypes: control (Pwiz), a moderate SMA model (C24), and a more severe model (X7C24). The *Smn* expression is significantly different in both disease models compared to control (Figure2). In the CNS dataset we only looked at two genotypes, the control (Pwiz) and the disease model (X7C24). In this set *Smn* expression was again significantly different between the disease model and control. We see *Smn* significantly differentially expressed in all our *Smn* RNAi models. We understand that there has already been much study in neuronal tissue along with promising therapeutics that target it; therefore, we chose to put more focus on muscle tissue for pathways that represent the best opportunities for further study.

	Paternal strain	Maternal strain	Genotype under analysis
PWIZ	P{UAS-WIXZ}	Mhc-Gal4	Mhc-Gal4 / PWIZ
C24	P{UAS-Smn$^{RNAi-C24}$}	Mhc-Gal4	Mhc-Gal4 /P{UAS-Smn$^{RNAi-C24}$}
X7C24	Df(3L)Smn^{X7},P{UAS-Smn$^{RNAi-C24}$}/TM6B	Mhc-Gal4	Mhc-Gal4 / Df(3L)Smn^{X7},P{UAS-Smn$^{RNAi-C24}$}

	Paternal strain	Maternal strain	Genotype under analysis
PWIZ	P{UAS-WIXZ}	P{Gal4-elav.L}3	P{Gal4-elav.L}3 / PWIZ
X7C24	Df(3L)Smn^{X7},P{UAS-Smn$^{RNAi-C24}$}/TM6B	P{Gal4-elav.L}3	P{Gal4-elav.L}3 / Df(3L)Smn^{X7},P{UAS-Smn$^{RNAi-C24}$}

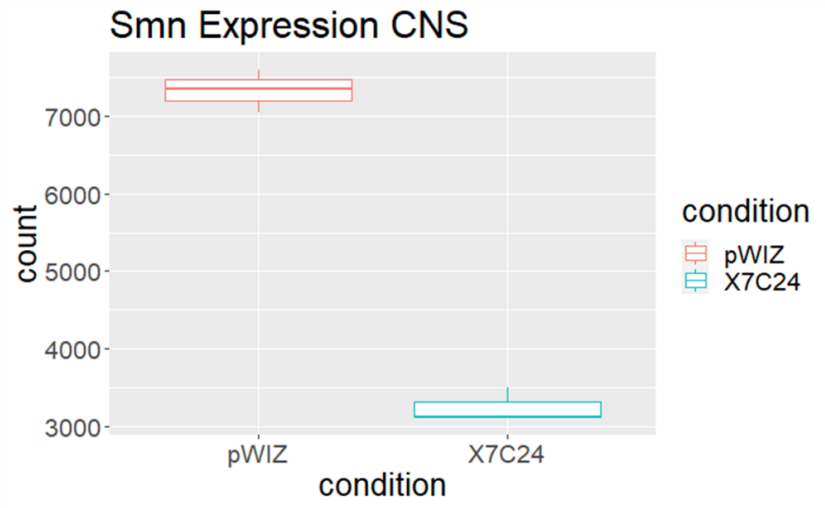


Figure 2. Muscle and CNS Crosses and *Smn* Expression

(Top) Mating scheme used to obtain animals with different levels of Smn expression and Smn expression levels in the three sample groups from muscle dataset. (Bottom) Mating scheme used to obtain animals with different levels of Smn expression and Smn expression levels in the two sample groups from CNS dataset.

We successfully generated lists of significantly differentially expressed genes with an adjusted p value of less than 0.05 in both the muscle and CNS datasets. We chose 0.05 as the cutoff both because it is conventional and because we wanted to ensure that we captured genes that were known to be impacted by *Smn* loss. In the muscle dataset we found 3693 genes were differentially expressed out of 16495 genes with a nonzero read count (Figure 3). Looking at lower p values we found 1076 genes had a p value less than 0.001 and 592 were lower than 0.00001. Of the significant genes 2201 were upregulated and 1492 were downregulated in X7C24 compared to control. Log2-fold changes ranged between -6.80 and 4.97. In the CNS dataset we found 998 genes were differentially expressed out of 16462 genes with a nonzero read count (Figure 3). Looking again at lower p values we found 452 genes had a p value less than 0.001, 296 with a p value less than 0.00001. In our statistically significant gene list 525 were found to be upregulated and 473 were downregulated compared to control. The log2 fold changes ranged between -4.30 and 3.28.

When these two lists of significant genes are compared, 319 genes overlap between them. This is a significant overlap with a p value of $4x10^{-17}$. Of these 319 overlap genes 233 (73 percent) of them change in the same direction meaning they are either both upregulated or both downregulated in both tissues.

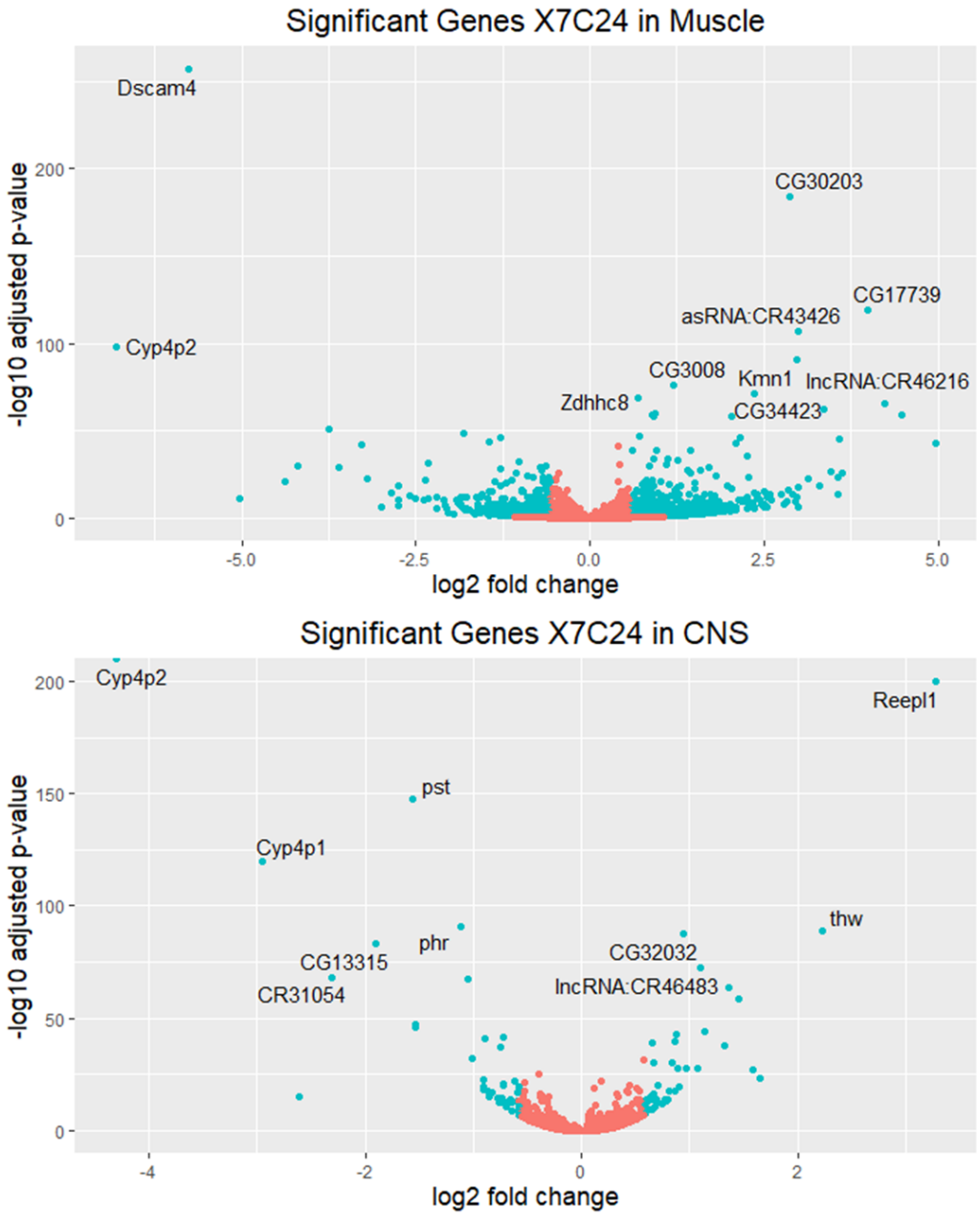


Figure 3. Volcano Plot of Significant Genes in Muscle and CNS

log10-adjusted p value along the y axis against the log2-fold change between X7C24 SMA disease model and control. Blue color designates a p value less than 0.05 and a fold change greater than 1.5. Positive x-values represent upregulation and negative x-values represent downregulation. Top 10 most significant genes labeled (top) muscle (bottom) CNS.

To compare our results to those previously established we compared these lists of significant genes to a list of 372 Smn genetic modifiers compiled from previous genome wide functional screens in *Drosophila* by Sen (Sen et al., 2013) and Chang (Chang et al., 2008). Our significant muscle gene list had 89 overlaps with this list, however, the overlap was not statistically significant at 0.09. The overlap with the significant CNS dataset was significant with a p value of 0.033 and an overlap of 30 genes. 14 genes overlapped in both the muscle and the CNS list with the genetic modifiers list.

We also took both control sample counts and put them through the DESeq2 pipeline to find genes not differentially expressed between the two controls, these genes were considered non-tissue-specific. We found 4,656 genes that were similarly expressed between the two controls. There is overlap between these non-tissue-specific genes and our lists of significant DE genes in both muscle and CNS. From our non-specific list 741 genes are changed only in muscle, 79 genes are changed only in CNS and 30 are changed in both.

Differential Exon Usage

We also looked for differential exon usage in our datasets given the popular hypothesis that SMN is required for multiple aspects of mRNA processing. Using DEXSeq and looking at results with a false discovery rate less than 0.05 we found 1909 significant exons (Figure 4) corresponding to 1109 genes in the muscle dataset. We found 1058 exons were upregulated and 851 were downregulated. When we analyzed the CNS data set we found 1022 significant exons (Figure 4) corresponding to 683 genes. Of these exons 529 were upregulated and 493 were downregulated. We found the overlap of these two significant gene lists to be 153 genes. This is a significant overlap and

represents 13.8% of the muscle genes and 22.4% of the CNS genes. Looking at the expression of the exons belonging to *Smn* we see significant differential usage in the muscle but not significant differential usage in the CNS. We compared these lists of DEU significant genes to the genetic modifiers list. Our DEU significant muscle gene list had 55 overlaps with this list and was statistically significant at a p value of 2.8×10^{-9}. The overlap with the DEU significant CNS gene list was also significant with a p value of 5.4×10^{-5} and an overlap of 31 genes. 10 genes were found in all three lists: the DEU muscle, DEU CNS, and genetic modifiers list.

We then compared our DE gene lists with our DEU gene lists. There were 338 genes that overlapped between both muscle datasets; this is a significant overlap. There were 134 genes that overlapped between the two CNS significant gene lists, again a statistically significant overlap. There are 12 genes that were found in all four gene lists.

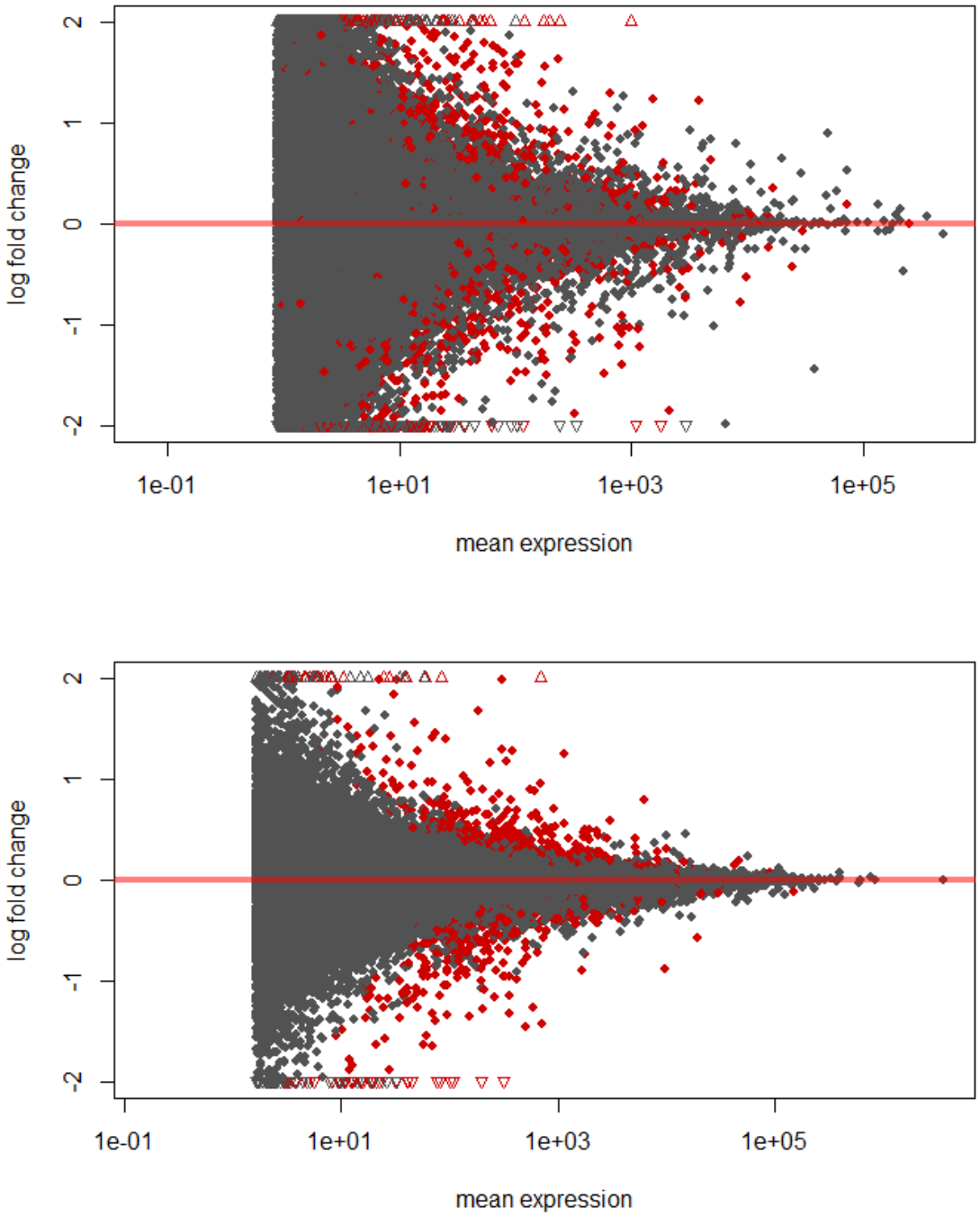
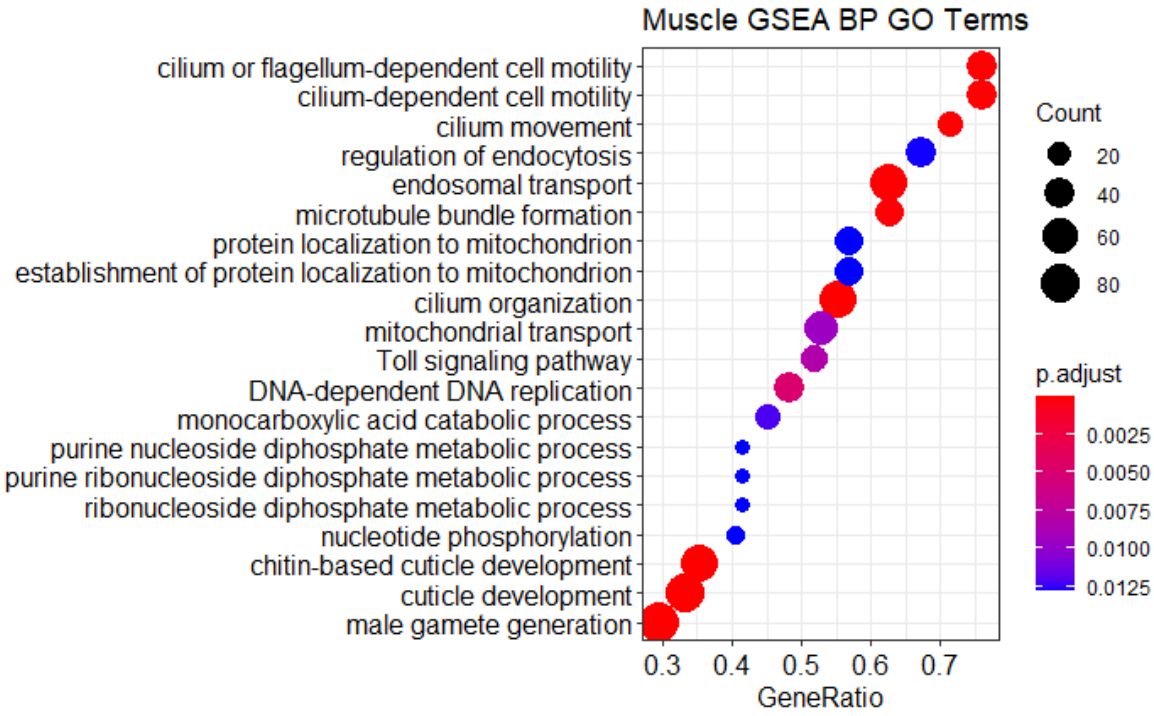


Figure 4. MA Plots of Muscle and CNS DEU

Log2-fold change between X7C24 and control against mean normalized count of exons. Red signifies the exon had a significant p value of less than 0.05. (Top) Muscle MA plot (Bottom) CNS MA plot.

Gene Set Enrichment Analysis

With our transcriptome analyzed by DESeq2 we now have log2-fold changes for each gene when compared between control and X7C24. With this we were able to perform Gene Set Enrichment Analysis (GSEA) on the whole DE dataset, looking at expression trends across all genes, not just the lists of ones that are statistically significant. The idea is that while single genes can have individual impacts on a pathway or process, smaller changes across a larger number of genes may combine to enact larger effects on whole pathways or processes that are only visible if the entire transcriptome is considered. If one only looks at statistically significant genes these overall affects could be missing from the analysis. Our muscle dataset has 32 statistically significant BP GO terms with a p value of 0.05 or less (Figure 5). Our CNS dataset has 20 significant BP GO terms with p values less than 0.05 (Figure 6). GSEA can also look at KEGG pathways: the muscle dataset has 7 significantly enriched pathways and the CNS dataset has 1 significantly enriched pathway as seen in Figure 7.



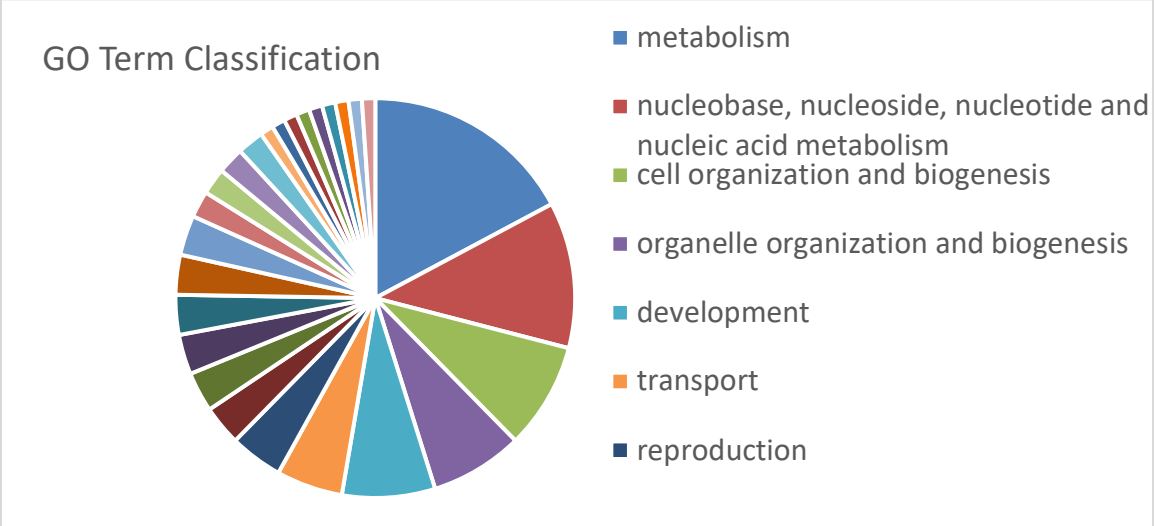


Figure 5. Significant GSEA Muscle GO BP Terms

(Top) Top 20 terms sorted by gene ratio (number of genes from the set that appear in that GO term list over the total number of genes in the go term list). P value designated by color gradient; number of genes associated with term designated by size of dot. (Bottom) Pie chart of GO term classification into Go_Slim2 Categories, largest categories labeled.

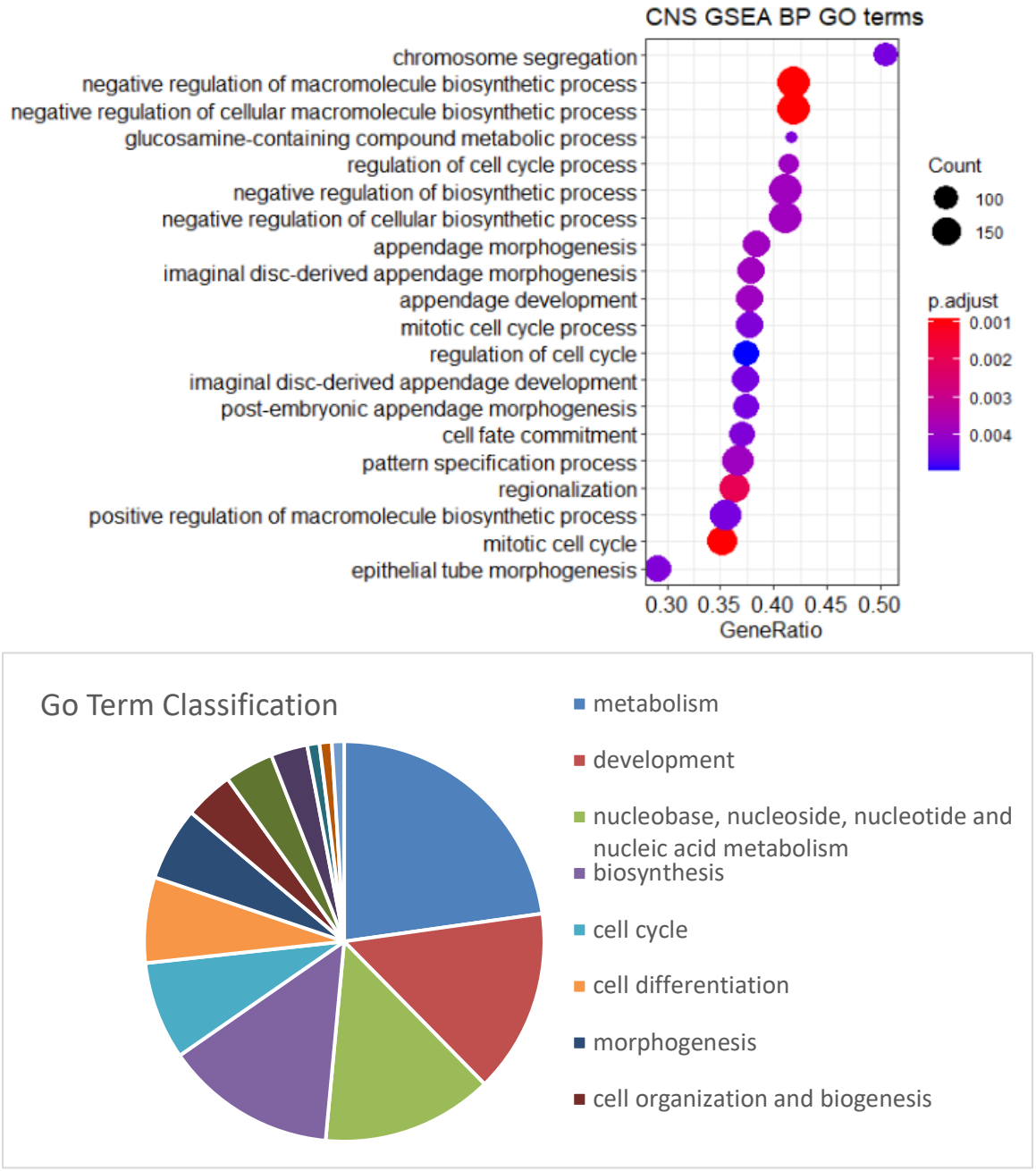


Figure 6. Significant GSEA CNS GO BP Terms

(Top) Terms sorted by gene ratio (number of genes from the list that appear in that GO term list over the total number of genes in the GO term list). P value designated by color gradient; number of genes associated with term designated by size of dot. (Bottom) Pie chart of GO term classification into Go_Slim2 Categories, largest categories labeled.

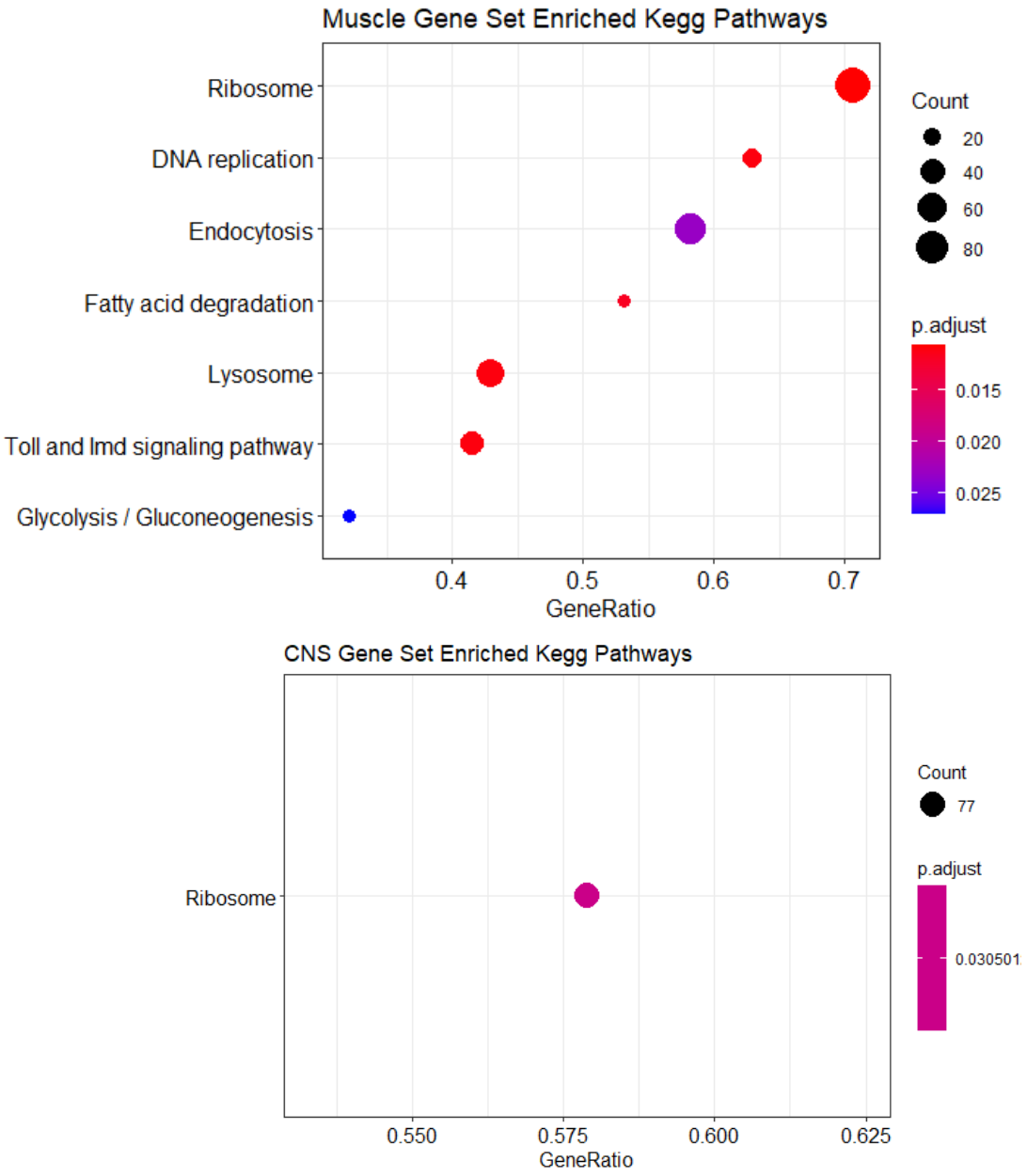


Figure 7. Significant GSEA KEGG Pathways in Muscle and CNS

The significantly enriched KEGG pathways using the fold changes of genes from the whole dataset. P value designated by color gradient; number of genes associated with the pathway designated by size of dot. Gene ratio is the number of genes from the gene list that appear in the KEGG pathway list over the total number of genes in the pathway list. (Top) Muscle. (Bottom) CNS.

Functional Analysis of Significant Genes

We also looked at the enriched GO terms of the significant genes in each list. Our DE significant gene list in the muscle found 11 enriched GO terms (Figure 8). Our DE significant gene list from CNS found 8 significantly enriched terms (Figure 9). The DEU muscle significant gene list had 80 significant results (Figure 10) and the DEU CNS list had 77 significant results (Figure 11). From our genetic modifiers list we found 61 significantly enriched terms (Figure 12).

We also looked at the enriched KEGG pathways of these lists. In the DE analysis we found 1 significant pathway from the muscle gene list and 7 in the CNS list (Figure 13). Neither the DEU nor the genetic modifiers list generated any significantly enriched pathways. However, when all the gene lists were put through the KEGG pathway mapper search function, over 100 pathways were found to have genes from each of our gene lists. In fact, most pathways were identified as hits from all four lists although the genes contributing to the pathways from each list were different.

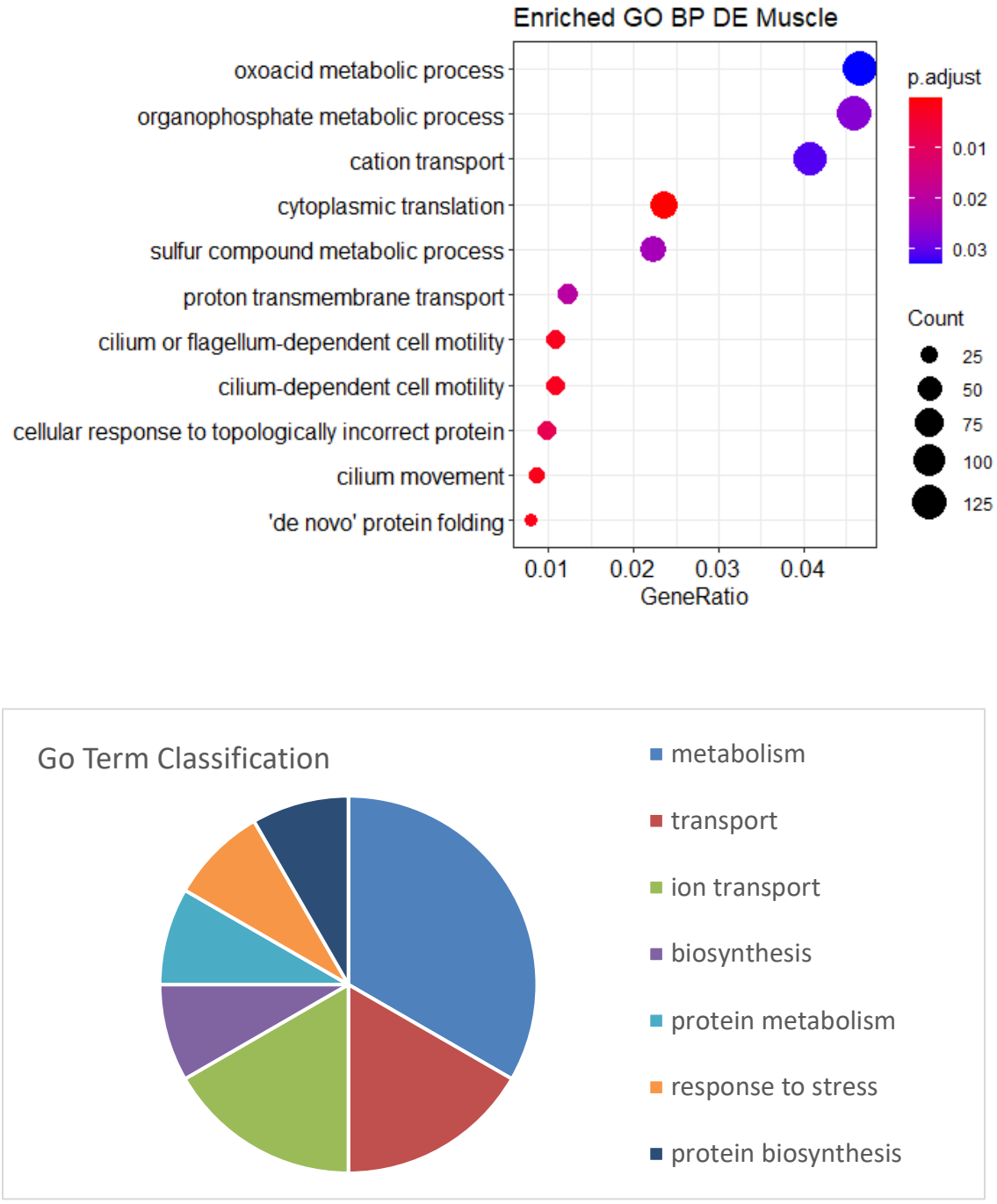


Figure 8. Enriched GO BP Terms from DE Significant Muscle List

(Top) GO terms sorted by gene ratio (number of significant genes related to GO term / total number of significant genes), adjusted p value indicated by color. Larger gene count is indicated by larger dot. (Bottom) Pie chart of GO term classification into Go_Slim2 categories, largest categories labeled.

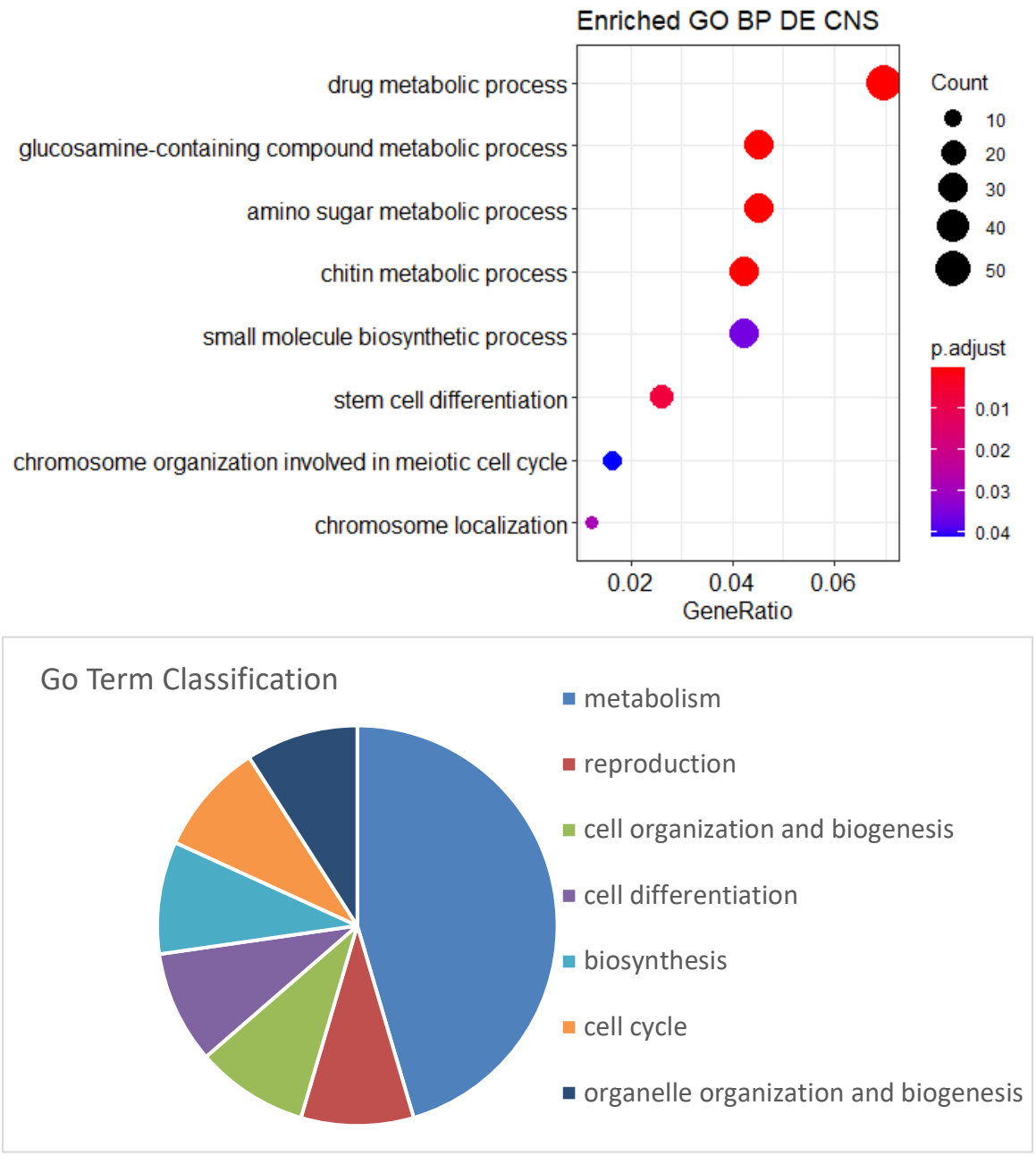


Figure 9. Enriched GO BP Terms from DE Significant CNS List

(Top) GO terms sorted by gene ratio (number of significant genes related to GO term / total number of significant genes), adjusted p value indicated by color. Larger gene count is indicated by larger dot. (Bottom) Pie chart of GO term classification into Go_Slim2 categories, largest categories labeled.

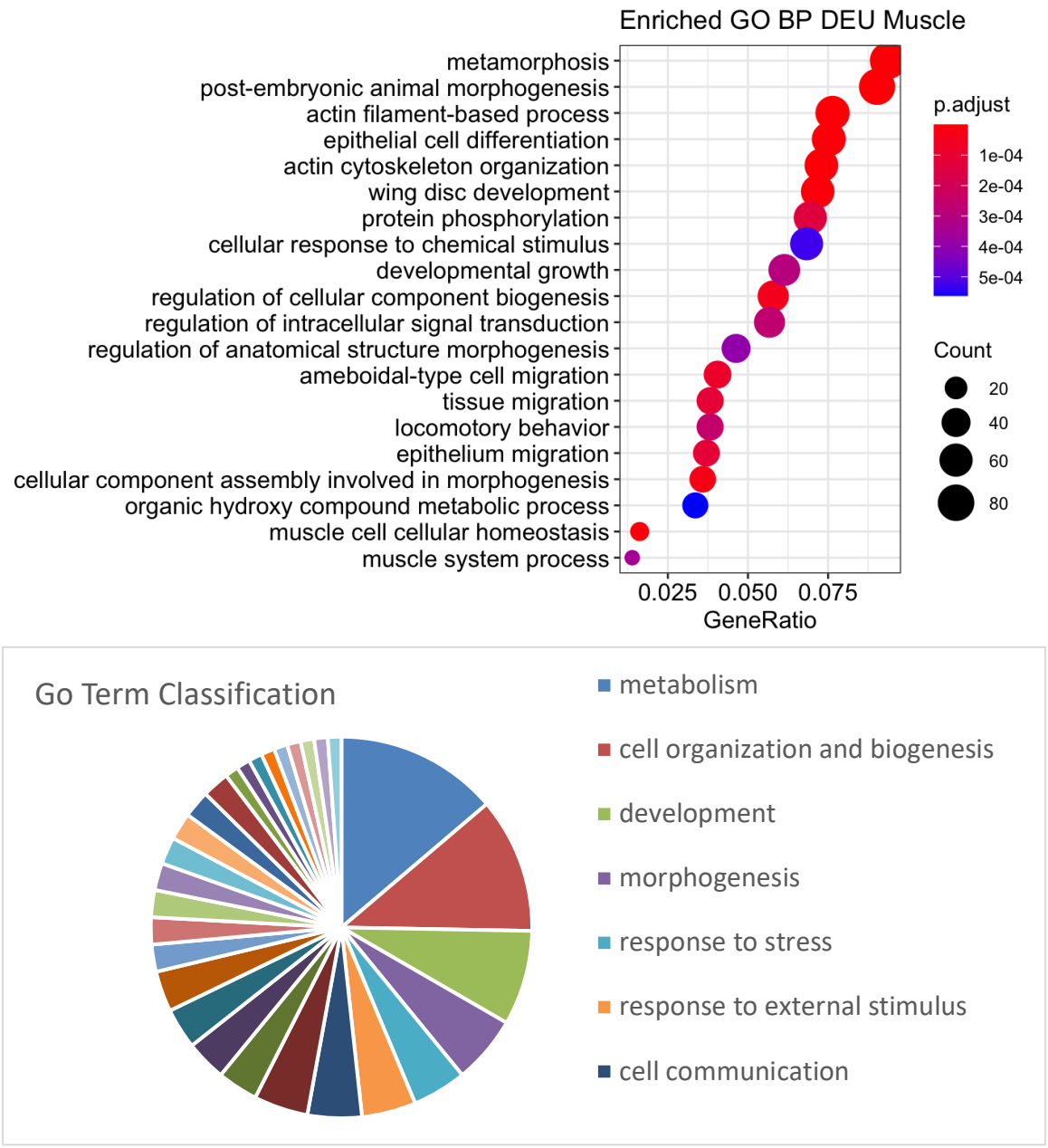


Figure 10.Top 20 Enriched GO BP Terms in DEU Significant Muscle List

(Top) Top 20 GO terms sorted by gene ratio (number of significant genes related to GO term / total number of significant genes); adjusted p value indicated by color. Larger gene count is indicated by larger dot. (Bottom) Pie chart of GO term classification into Go_Slim2 categories, largest categories labeled.

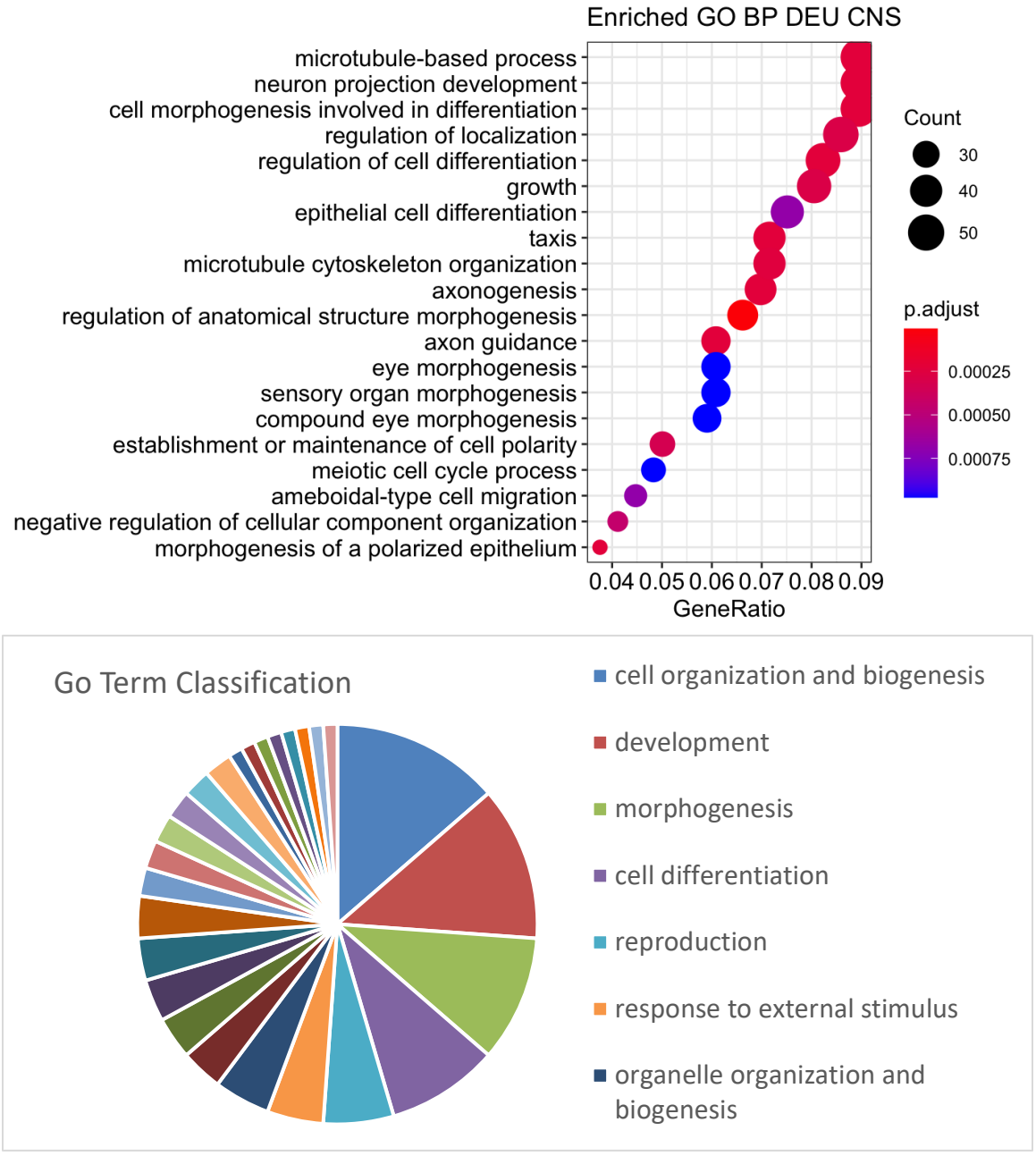


Figure 11.Top 20 Enriched GO BP Terms in DEU Significant CNS List

(Top) Top 20 GO terms sorted by gene ratio (number of significant genes related to GO term / total number of significant genes); adjusted p value indicated by color. Larger gene count is indicated by larger dot. (Bottom) Pie chart of GO term classification into Go_Slim2 categories, largest categories labeled.

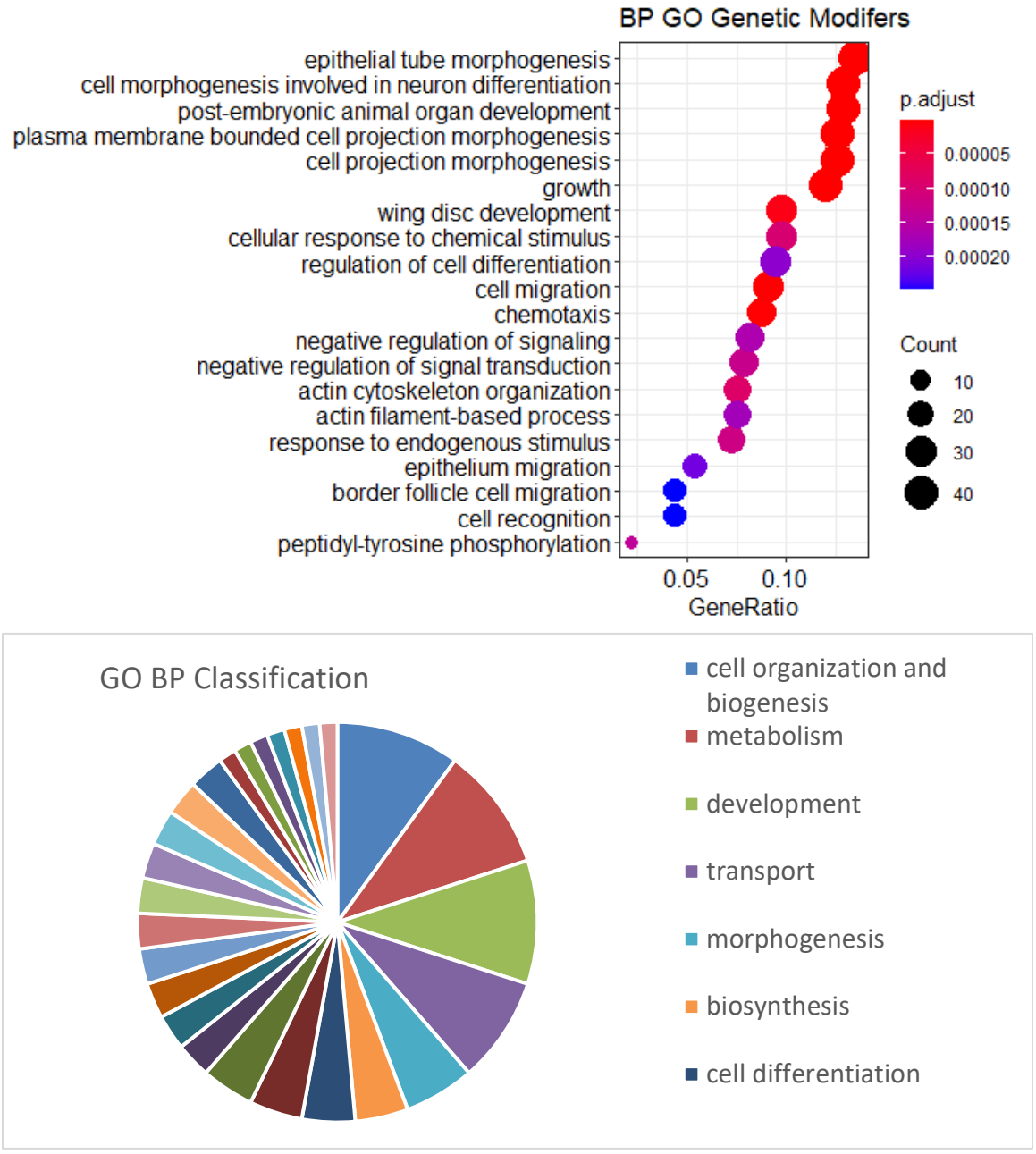


Figure 12.Enriched GO BP terms in Genetic Modifiers list

(Top) Top 20 GO terms sorted by gene ratio (number of significant genes related to GO term / total number of significant genes); adjusted p value indicated by color. Larger gene count is indicated by larger dot. (Bottom) Pie chart of GO term classification into Go_Slim2 categories, largest categories labeled.

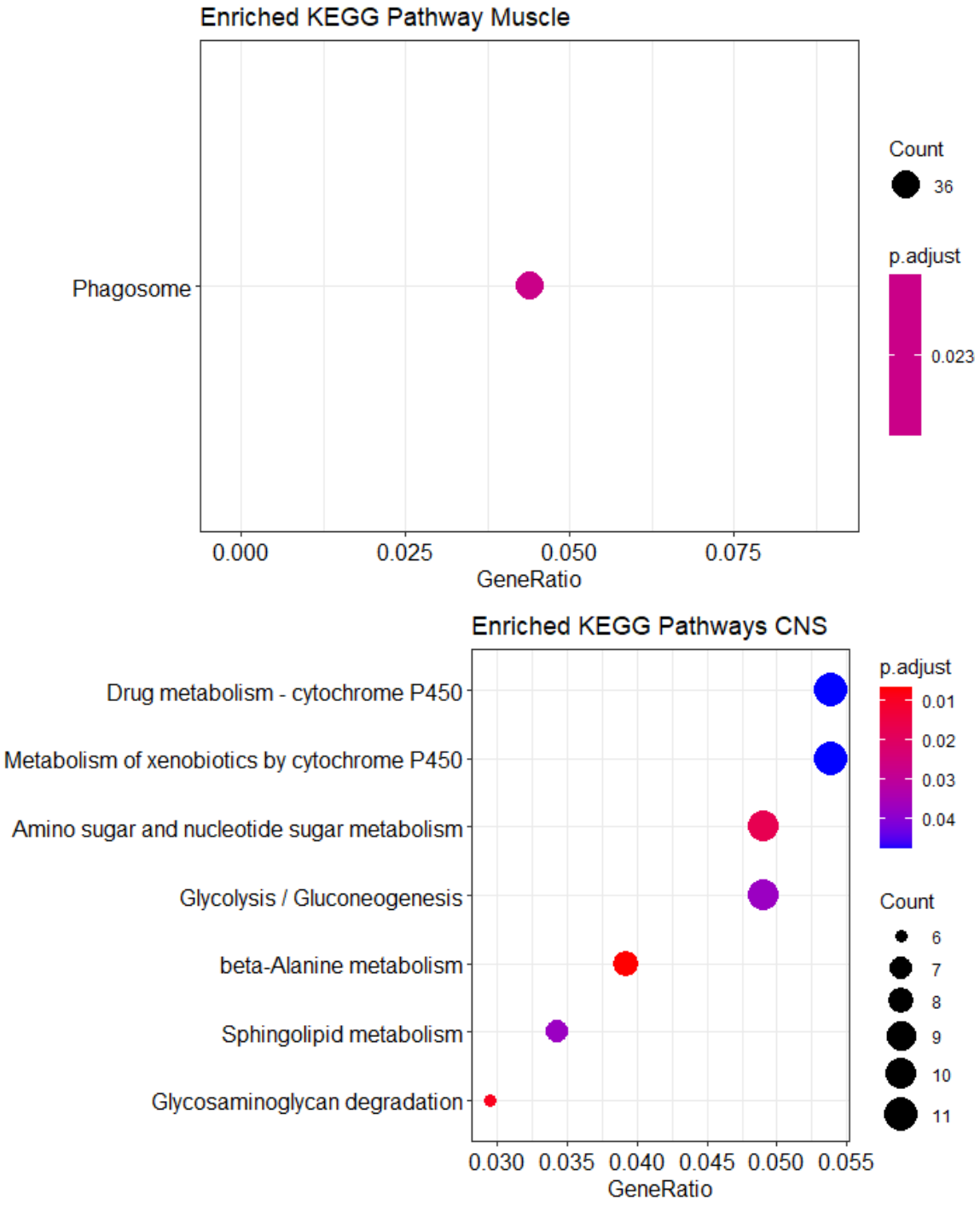


Figure 13.Enriched KEGG Results in DE Significant Muscle and CNS Lists

Significantly enriched KEGG pathway results. Terms sorted by gene ratio. P value designated by color gradient; number of genes associated with term designated by size of dot. (Top) Muscle (Bottom) CNS

Dose-dependent Filtering

Our significant muscle gene list is very large and has potential compensatory gene expression that is unrelated to *Smn* loss of function. In SMA, the severity of the disease is determined by the amount of SMN available to the cell. Our muscle data included two disease models, one more severe than the other. In the more severe model less *Smn* is expressed and the phenotype of the disease is worse. Our CNS dataset only included one disease model as it is more often studied and we wanted to focus on the opportunities in muscle. Using these two models of different severity, we are also able to see what genes are affected in a similar dose-dependent manner by looking at the normalized counts and finding genes in the significant gene list that follow the pattern Pwiz < C24 < X7C24 or X7C24 < C24 < Pwiz. In this way we were able to subset the list of significant (adjusted p value less than 0.05) genes in muscle into those that are dose-dependent on *Smn*. We believe that doing this will help to limit our significant muscle gene list to those genes that are most likely to be associated with *Smn* and involved in the SMA phenotype. We found 1865 genes that fit this dose-dependent pattern. When we compare this to the CNS list, in which we did not determine dose dependence, we see an overlap of 185 genes. This overlap is significant with a p value of 6.71×10^{-15}. Looking at the overlap between the two lists, 140 of the 185 genes are differentially expressed in the same direction, meaning 76% of them are upregulated or downregulated in both tissues. We compare this filtered list to the genetic modifiers list and find 60 genes overlap, and the p value is 5.4×10^{-4}. Where the unfiltered muscle list had not been significant, this dose-dependent list is.

We were able to apply the same dose-dependent filtering to the DEU exons and from that extrapolate the affected genes. We found 897 exons that fit the pattern based on their exon usage coefficients. These exons correspond to 581 genes. Comparing this list to the genetic modifiers list finds 35 genes overlap, a significant overlap with a p value of 2.3×10^{-8}. There are 114 genes that are found in both DE and DEU dose-dependent lists. This overlap is statistically significant with a p value of 3.9×10^{-11}.

While the exact enriched GO terms vary between the unfiltered and filtered gene lists, the larger classifications of GO terms are mostly the same before and after filtering for dose dependency. These GO terms are seen in Figure 14 for and 15 for DE and DEU sets respectively. There are 13 DE dose dependent enriched muscle BP GO terms and 47 DEU dose dependent enriched muscle BR GO terms.

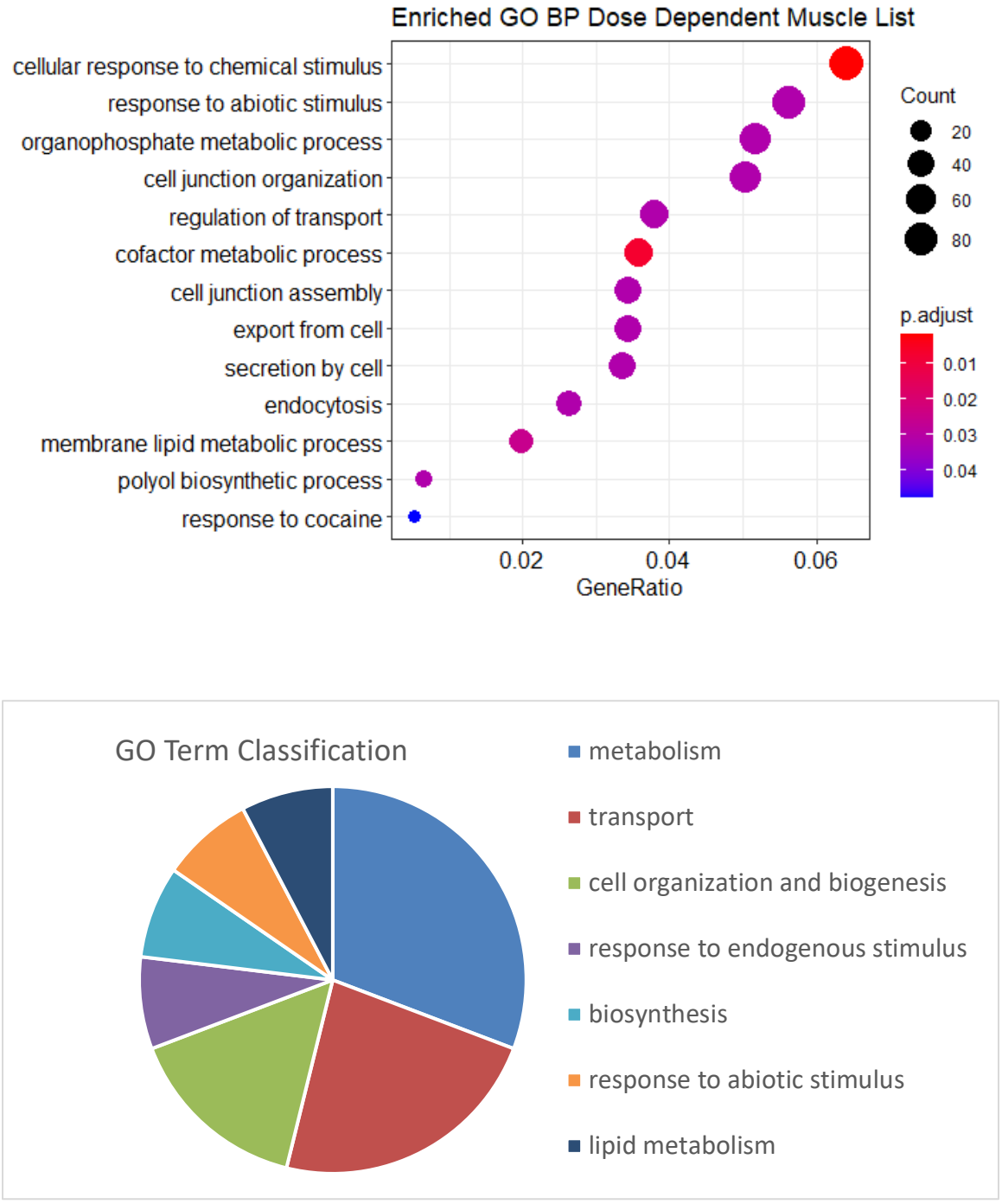


Figure 14.Enriched GO BP Terms from DE Muscle Dose-dependent List

(Top) GO terms sorted by gene ratio (number of significant genes related to GO term / total number of significant genes), adjusted p value indicated by color. Larger gene count is indicated by larger dot. (Bottom) Pie chart of GO term classification into Go_Slim2 categories, largest categories labeled.

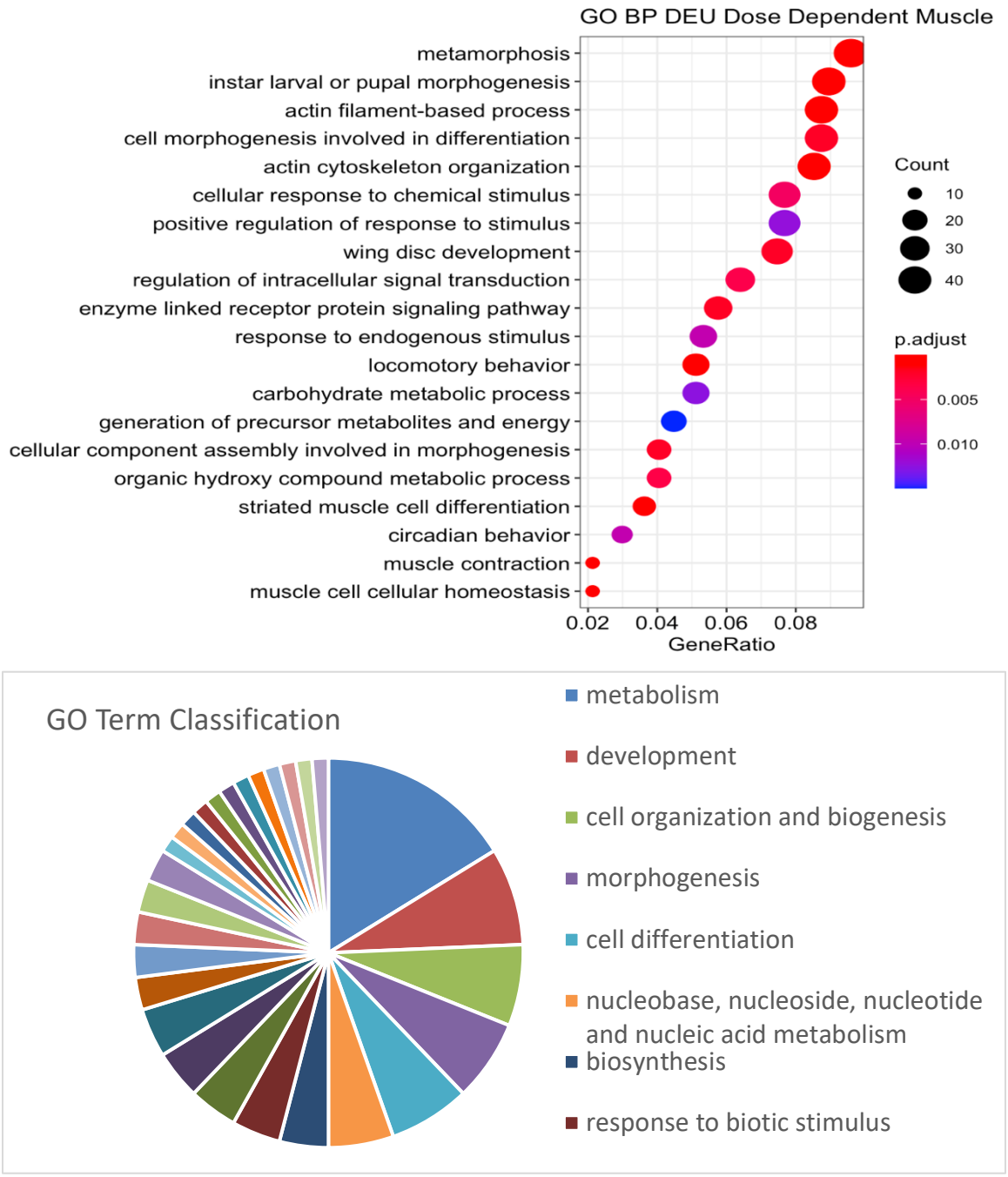


Figure 15.Enriched GO BP Terms from DEU Muscle Dose-dependent List

(Top) GO terms sorted by gene ratio (number of significant genes related to GO term / total number of significant genes), adjusted p value indicated by color. Larger gene count is indicated by larger dot. (Bottom) Pie chart of GO term classification into Go_Slim2 categories, largest categories labeled.

Functional Analysis of Filtered Gene Lists

We further explored our gene lists by filtering out the genes that overlapped between both lists. Once the overlapping genes were removed from the muscle dataset, we found 8 enriched GO BP terms (Figure 16). The unique muscle list was also filtered for dose dependence. These genes were analyzed for BP GO terms and then simplified and we found that 16 GO terms were significantly enriched with a p value of less than 0.05 (Figure 17). We also looked at the enriched BP GO terms in the CNS data, we separated out genes that appeared in the significant muscle dataset to make a list of CNS tissue-specific significant genes. We found 15 significantly enriched GO BP terms (Figure 18). We applied the same filtering to the DEU gene sets for dose dependence and for tissue-specificity. We found 57 GO BP terms in unique muscle (Top 20 shown in figure 19), 42 significant BP GO terms in unique dose-dependent muscle (Top 20 shown in Figure 20) and 54 significant GP BP terms in unique CNS (Top 20 shown in figure 21). We also looked for enriched KEGG pathways in our tissue-specific gene lists. However, despite having many genes appearing in many different pathways they did not generate any significantly enriched pathways.

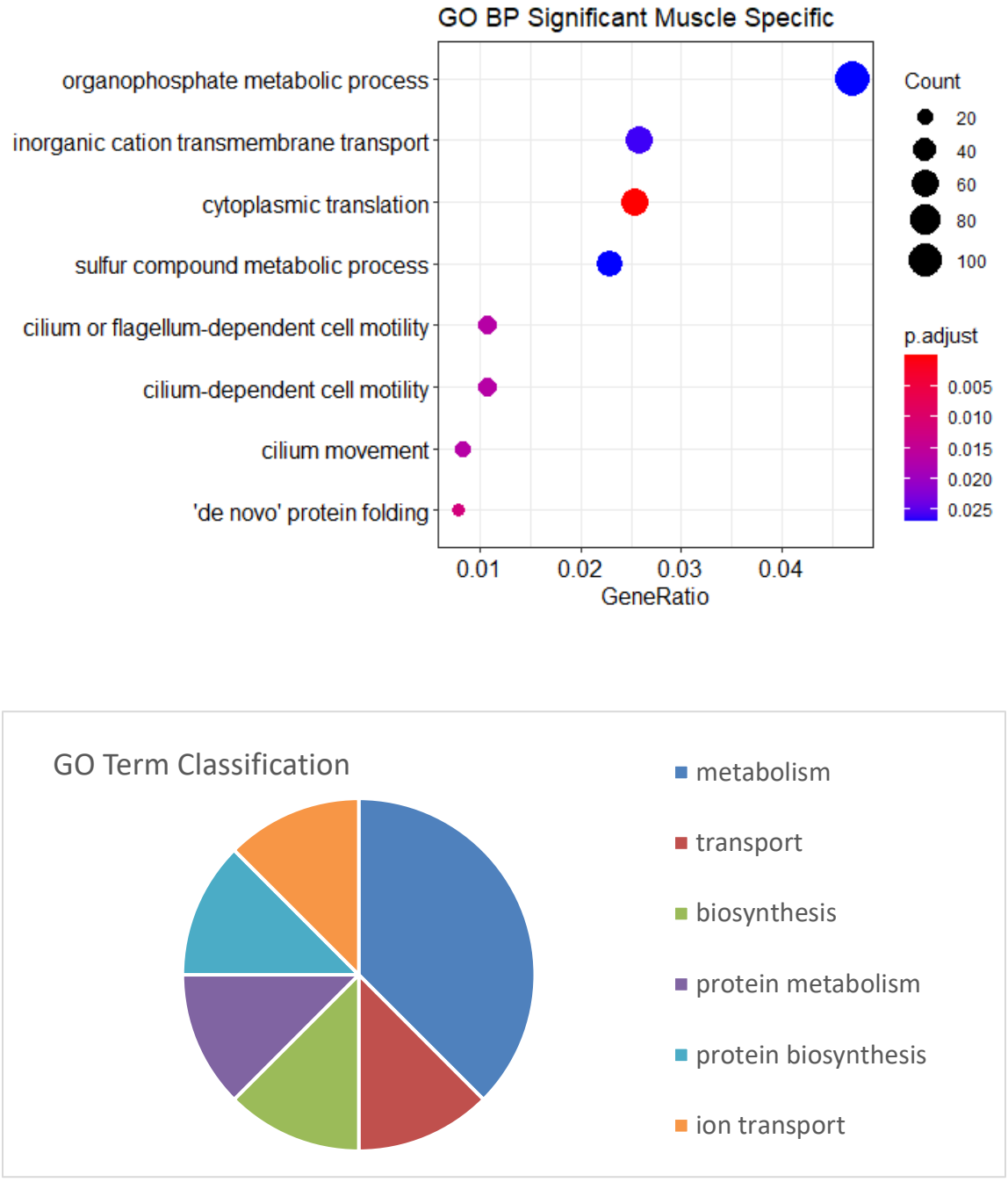
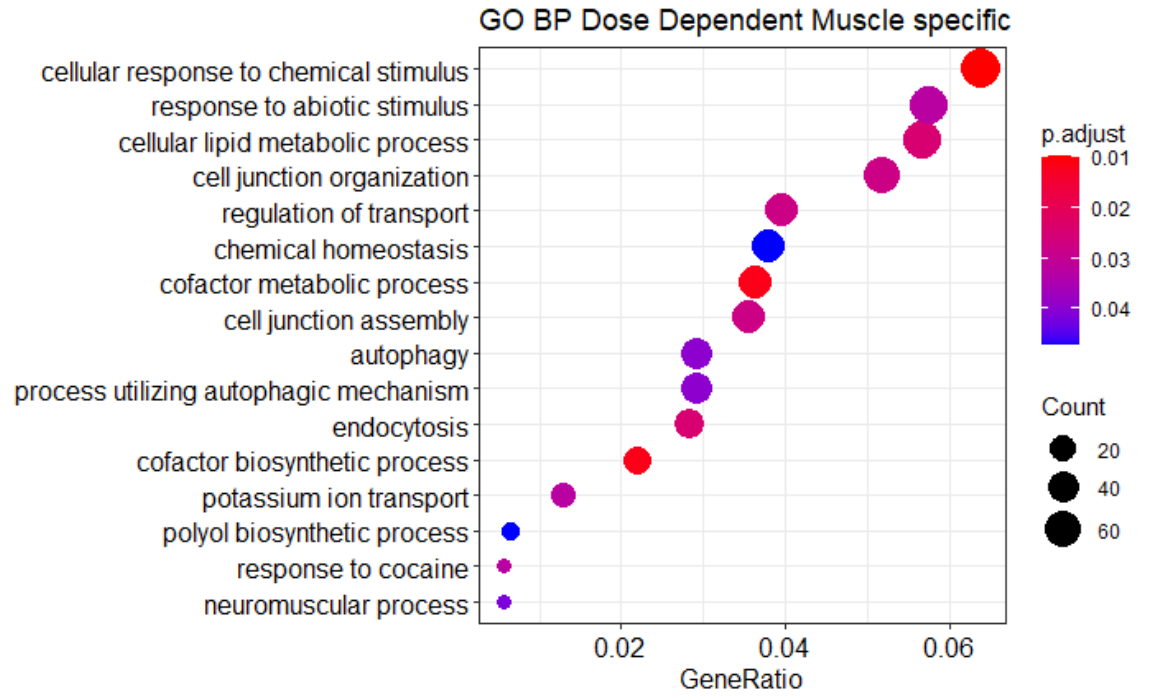


Figure 16.Enriched GO BP Terms from Significant Muscle Specific Gene List

(Top) GO BP terms sorted by gene ratio (number of significant genes related to GO term / total number of significant genes), adjusted p value indicated by color. Larger gene count is indicated by larger dot. (Bottom) Pie chart of GO term classification into Go_Slim2 categories, largest categories labeled.



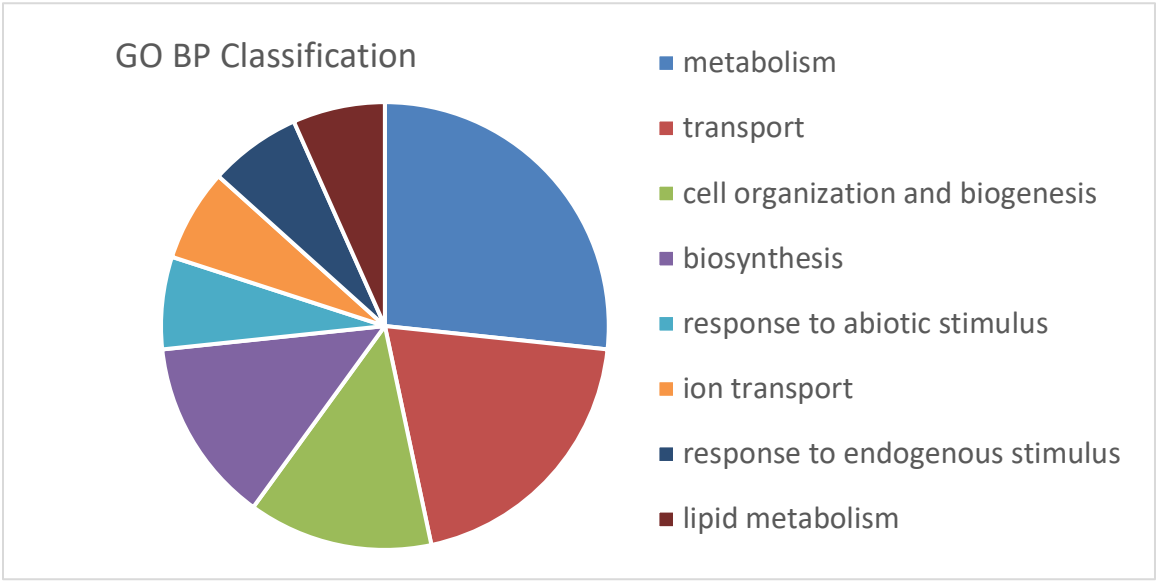
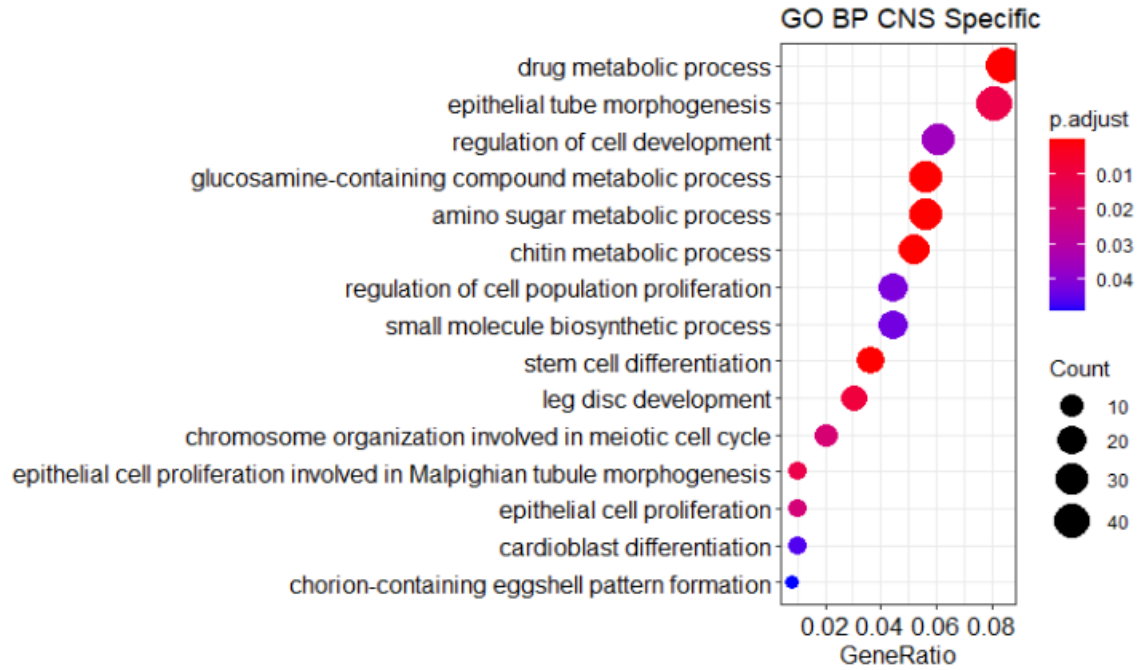


Figure 17.Enriched GO BP Terms from Dose-Dependent Muscle Specific List

(Top) GO BP terms sorted by gene ratio (number of significant genes related to GO term / total number of significant genes), adjusted p value indicated by color. Larger gene count is indicated by larger dot. (Bottom) Pie chart of GO term classification into Go_Slim2 categories, largest categories labeled.



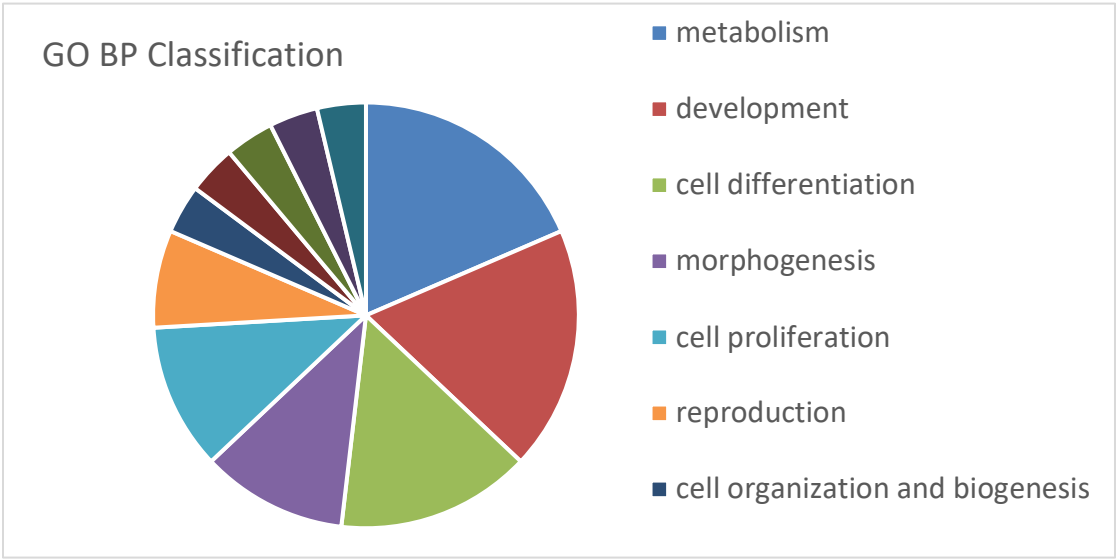


Figure 18.Enriched GO BP Terms from Significant CNS Specific List

(Top) GO BP terms sorted by gene ratio (number of significant genes related to GO term / total number of significant genes), adjusted p value indicated by color. Larger gene count is indicated by larger dot. (Bottom) Pie chart of GO term classification into Go_Slim2 categories, largest categories labeled.

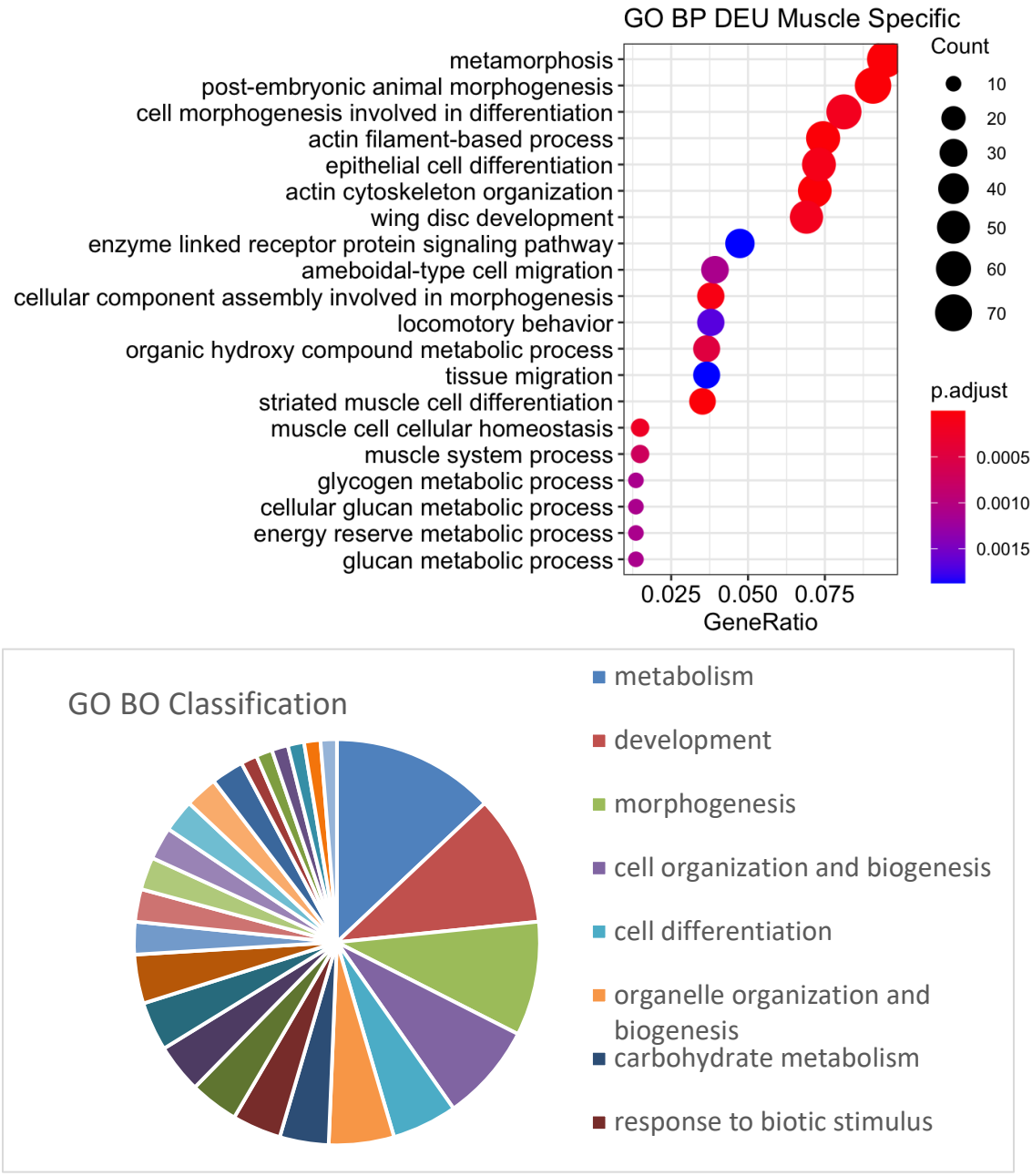


Figure 19.Top 20 GO BP Terms from DEU Muscle Specific List

(Top) Top 20 GO BP terms sorted by gene ratio (number of significant genes related to GO term / total number of significant genes); adjusted p value indicated by color. Larger gene count is indicated by larger dot. (Bottom) Pie chart of GO term classification into Go_Slim2 categories, largest categories labeled.

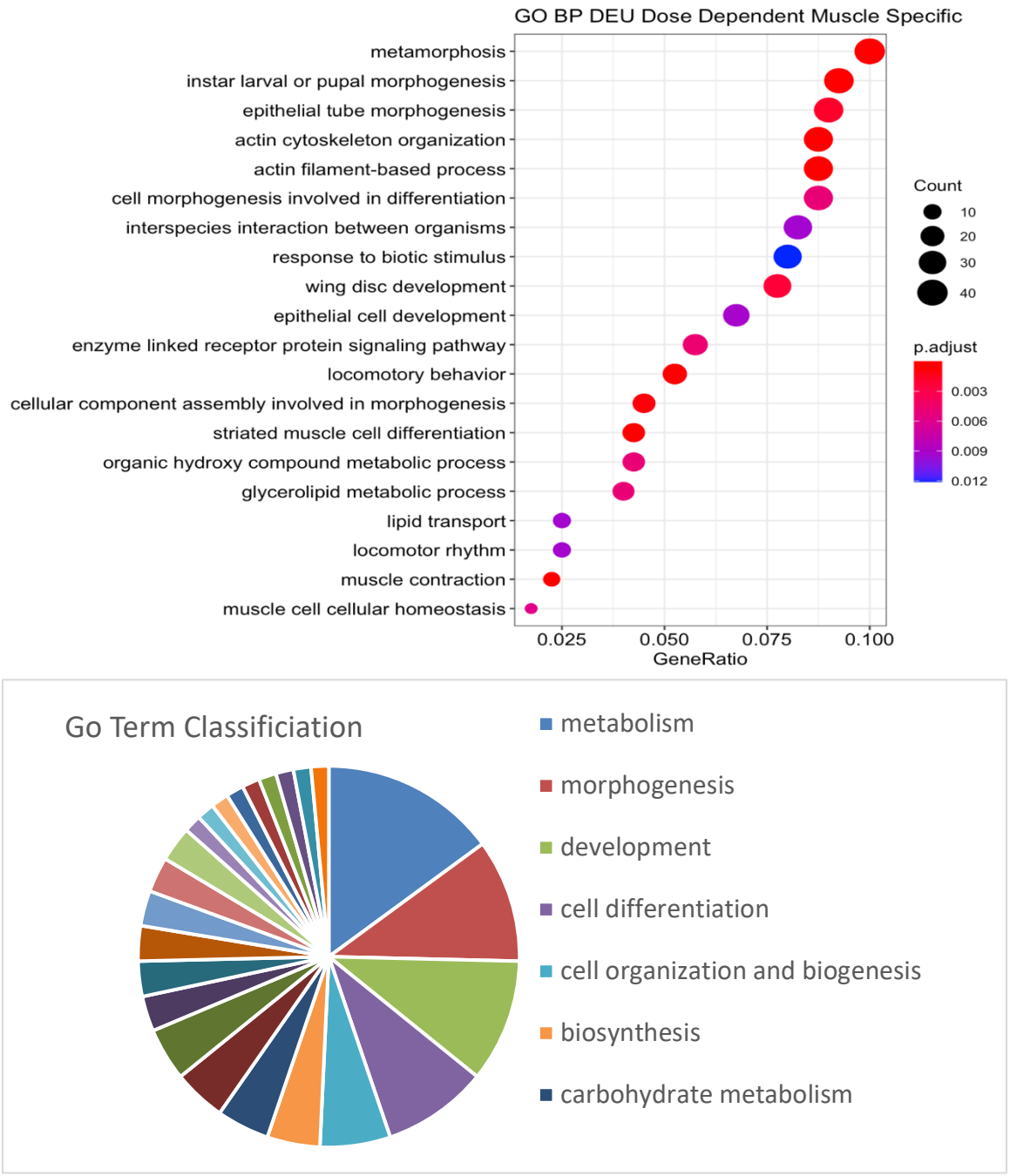


Figure 20.Top 20 GO BP Terms from DEU Dose-dependent Muscle Specific List

(Top) Top 20 GO BP terms sorted by gene ratio (number of significant genes related to GO term / total number of significant genes); adjusted p value indicated by color. Larger gene count is indicated by larger dot. (Bottom) Pie chart of GO term classification into Go_Slim2 categories, largest categories labeled.

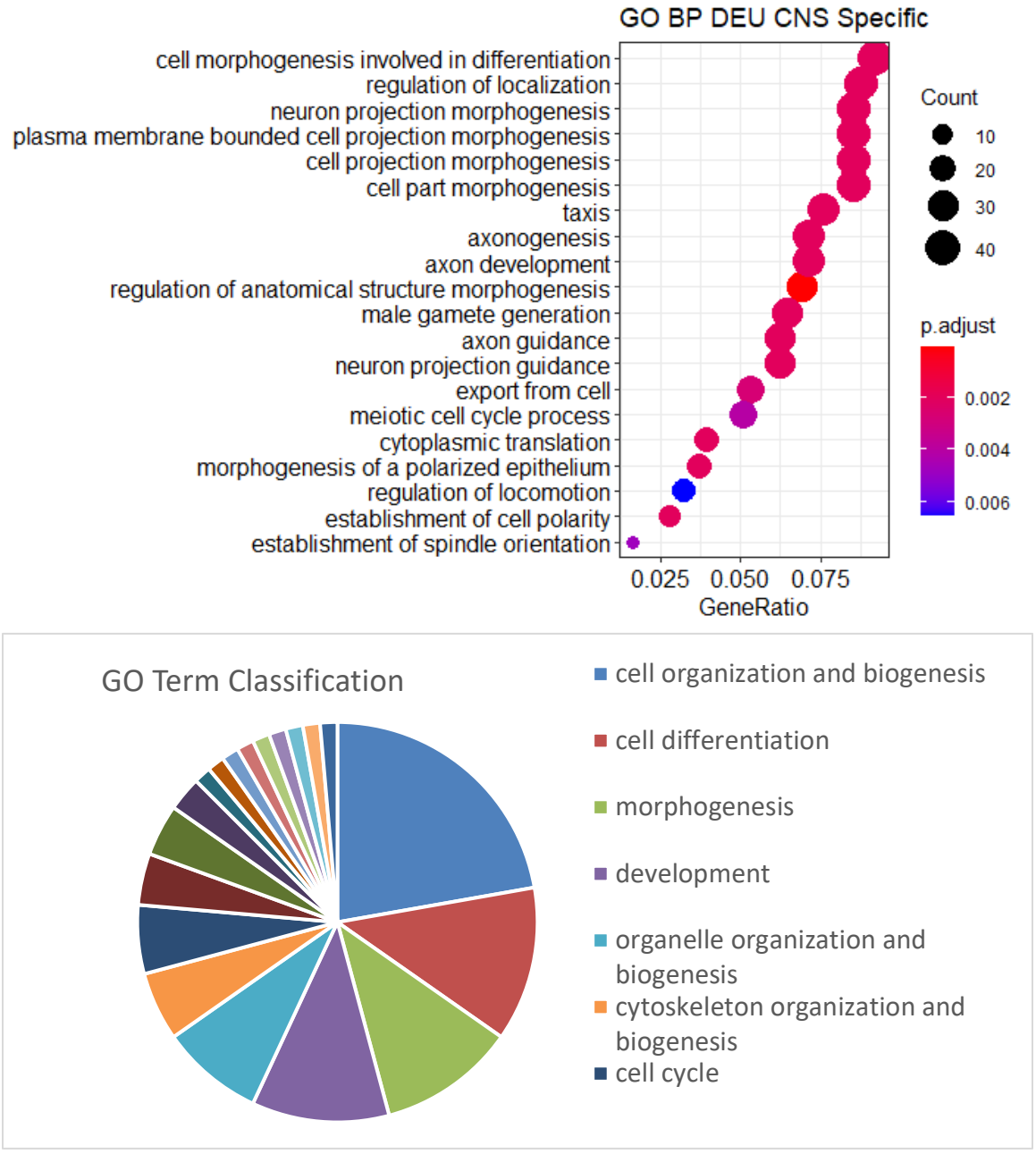


Figure 21.Top 20 GO BP Terms from DEU CNS Specific List

(Top) Top 20 GO BP terms sorted by gene ratio (number of significant genes related to GO term / total number of significant genes); adjusted p value indicated by color. Larger gene count is indicated by larger dot. (Bottom) Pie chart of GO term classification into Go_Slim2 categories, largest categories labeled.

Developmental Arrest Investigation

Other investigations of *Drosophila melanogaster Smn* loss have shown developmental arrest (Garcia et al., 2013). Therefore, we compared our lists of significantly differentially expressed genes to lists of stage specific genes. These lists were generated from the Flybase tool "Search RNA-Seq data by expression profile" using modENCODE expression data. Three different metrics were used. First, we searched for genes that were L1 or L2 stage specific by finding genes with expression patterns that were expressed in L1, L2, or both L1 and L2 at least at a "moderate" level and expressed "none or extremely low level" in all L3 stages. We found 10 genes that fit this stage specific profile. There was no significant overlap with our DE and DEU significant genes; only one gene overlapped with the DE muscle dataset and none with the CNS dataset. The one overlapping gene is overexpressed in X7C24 compared to control. Second, we searched for genes that were expressed in L1, L2, or both L1 and L2 at least at a "moderate" level and expressed at most "very low" in all L3 stages and found 84 genes that fit this stage specific profile. This larger less strict list found 15 overlaps with the DE muscle list, 1 with the DE CNS list, and 1 with both DE lists; however, this overlap again was not statistically significant. In the DEU gene lists we found an overlap of 8 in the muscle DEU genes and 3 in the CNS DEU genes but neither overlap was statistically significant. Last, we looked for genes that were L3 stage specific by looking for ones expressed at least at "moderate" level in L3 and "none or extremely low" in L1 and L2 instar stage. We found 15 genes that fit this expression profile on Flybase. We saw 1 gene overlap with the DE muscle list, 1 with the DEU muscle list, 2 gene overlap with the CNS list and none with the CNS DEU list. These overlaps are not statistically

significant. If our larva were in developmental arrest we would expect to see a significant number of these L3 genes underexpressed in our *Smn* mutant and a significant amount of the L1 and L2 specific genes overexpressed.

DIOPT Analysis

We took our lists of significant genes and looked for conserved orthologs between fly and human. From our list of 3,693 significant genes in muscle 2,648 genes (72%) generated hits corresponding to 10,300 possible human orthologs. Filtering out all the low-ranking hits we found 2,409 genes with their best scoring orthologs achieving a high or moderate rank. Out of our significant CNS gene list of 998 genes we found 709 genes (71%) had possible human orthologs, matching to 3,384 ortholog hits. When the low-ranking hits are filtered out there were 663 gene hits with their best orthologs at either high or moderate ranking. We also ran the list of DEU genes through the DIOPT tool. Out of our list of 1,109 muscle genes we found 912 (82%) had possible corresponding human orthologs. Of these genes, 839 had their highest match ranked as high or moderate. The total human orthologs found that corresponded to the DEU muscle genes were 3,910 total hits. Out of the list of 682 CNS DEU genes DIOPT found that 560 genes (82%) had possible orthologs, 515 of these were ranked as high or moderate. The total number of orthologs matches found included 2,397 human genes for the DEU CNS list. Limiting our significant gene list to just those conserved genes increases the number of significant GO BP terms. It does not change the number of significant KEGG pathways.

Chapter IV

Discussion

Conclusions

Despite the Survival of Motor Neuron protein being conserved and vital for survival in many species, its full role in the cell is still not fully understood (Groen et al., 2018). It appears to touch many different pathways and processes in different ways. SMN is ubiquitous in all cell types (Lefebvre et al., 1997) and its complete loss is embryonic lethal. It is necessary for cellular survival and function (Schrank et al., 1997). This contrasts the SMA disease phenotype in that the most pronounced and striking symptoms of the disease are seen explicitly in motor neurons and the neuromuscular system (Kolb & Kissel, 2011). Furthermore, multiple studies have shown that SMA relevant phenotypes can be elicited by knocking out SMN in one individual tissue at a time: specifically in muscle or in neurons (Kim et al., 2020; Laird et al., 2016). We know SMN is needed in both these tissues for normal cell function, as it is needed in all cells, but its functional impact on these two tissues was the central question for our investigation.

The first major conclusion we can draw from our findings is that *Smn* exhibits a significant degree of tissue-specificity in function. In our DE and DEU results we find a statistically significant overlap between *Smn*-dependent gene expression in muscle and CNS. However, most of the altered gene expression is found uniquely in one tissue or the other. One expects to see different expression profiles in different cell types, as each

tissue has a number of distinct functions with different underlying molecular mechanisms. Our two normal control genotypes used for each tissue have the same empty vector driven by tissue-specific GAL4 drivers, however, these RNAs were extracted from two non-overlapping tissue samples (larval ventral nerve cord and central brain complex versus larval body-wall muscle pelt). While there could be some differences in gene expression between these control genotypes due to the tissue-specific expression of Gal4, both genotypes drive moderately high levels of this exogenous transcription factor that lacks natural target sequences in the *Drosophila* genome (Duffy, 2002). Thus, the major difference between our two control sample types is the tissue that they came from. In order to identify genes that are expressed in common at similar levels in both CNS and muscle pelt, we ran our two control datasets through the DESeq2 pipeline to compare. We obtained a list of 4,656 genes that are not differentially expressed between wild type CNS and muscle pelt. This list was then compared to our two lists of genes that are significantly changed when control and *Smn* knockdown is compared in a given tissue, in order to ask if any of these tissue-non-specific genes are differentially affected by *Smn* knockdown in CNS or muscle, seen in Figure 22.

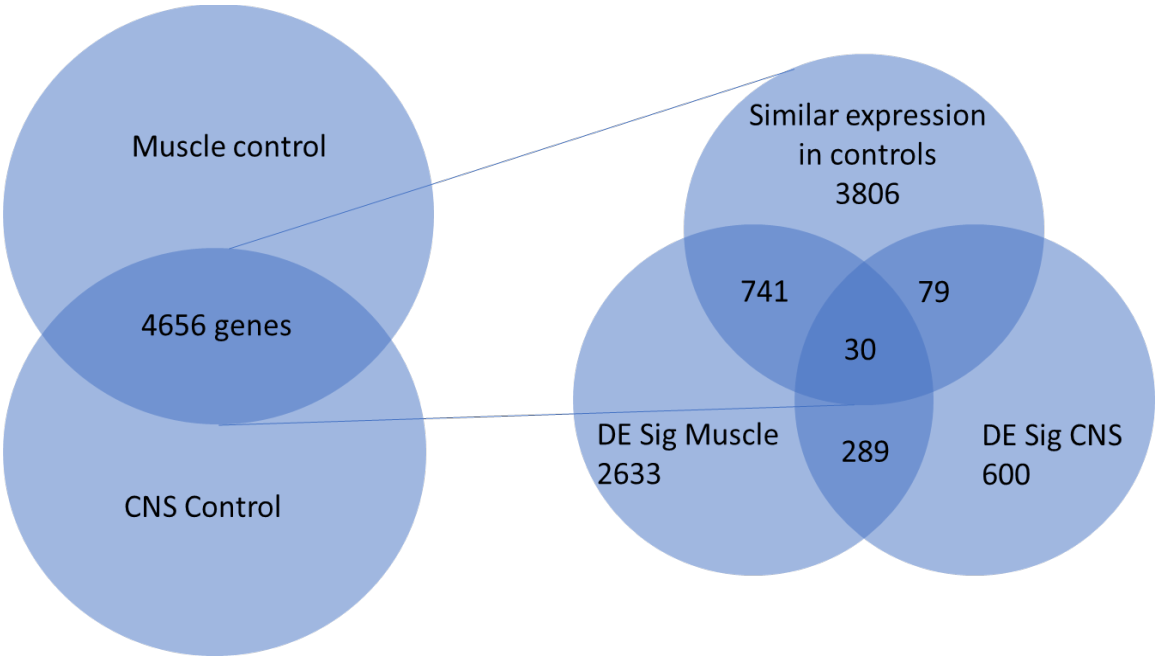


Figure 22.Non-Tissue-Specific Genes Change Expression Level in *Smn* Knockdown

When comparing the two control datasets, 4,656 genes are expressed at similar levels between tissues. Comparing these non-tissue specific genes to our DE genes from muscle and CNS (Genes that differ between Smn knockdown and control).

Interestingly, among genes that are normally expressed at the same levels in CNS and muscle, we find that a number of genes exhibit tissue-specific sensitivity to *Smn* function: 741 genes are changed only in muscle, 79 genes are changed only in CNS, and 30 genes are changed in both (Figure 22). This suggests that *Smn* displays a degree of tissue-selective regulatory function for a subset of downstream genes. While there appear to be genes that are altered in both presynaptic and postsynaptic cells, consistent with the known ubiquitous aspects of *Smn* function, the subset of cell type specific genes and pathways revealed in our RNA-Seq datasets suggest that some downstream

mechanisms will be altered only in neurons or muscle. This raised the question of what functional pathways might be particularly sensitive to *Smn* activity in neurons or muscle.

We next explored the tissue-specificity of *Smn*-dependence at the level of downstream biological processes and pathways. When we looked at the significantly enriched GO and KEGG terms found among *Smn*-dependent genes in each tissue, we found that different biological processes were represented in every comparison between muscle and CNS. The vast majority of terms were tissue-specific. Moreover, the small group of overlapping GO terms are not the most significant GO terms in each tissue's list. In the GSEA lists we barely found any overlap in GO terms between the two lists; only 4 terms: DNA replication, DNA-dependent DNA replication, imaginal disc morphogenesis, and post-embryonic animal organ morphogenesis. The GSEA-significant KEGG pathway terms only have one overlap, Ribosome, the rest are specific to muscle. When we looked at the GO enrichment analysis of the DE significant gene lists there was no overlap of GO BP terms between muscle and CNS. The significant KEGG pathways from the DE gene lists displayed no overlap either. These enriched KEGG pathways are statistically significant based on analysis using a hypergeometric equation statistical method that determines whether genes from annotated KEGG pathways are present more than would be expected in our list of significant DE genes. Many more pathways were not statistically significant but still contained some of our genes of interest. In these non-significant pathways we found almost full overlap between muscle and CNS, however the specific genes in the pathways often differed in each tissue. Suggesting that while SMN may interact with the same pathways in different tissues, the specific pathway nodes are tissue specific. Comparable to our DE dataset results, 74% of the DEU GO

terms from each tissue were unique to that tissue. While we did not get any significant KEGG terms from these DEU gene lists we found many non-significant pathways are touched by genes from our DEU gene lists. These pathways generally overlapped but represented different genes contributing to the pathway from each list. We also compared the enriched GO terms that are generated when the overlapping genes are removed from each list. These gene lists are the tissue-specific gene lists. There is no overlap in the enriched GO BP terms generated from the DE tissue-specific muscle and tissue-specific CNS gene lists. Similarly, over 90% of the enriched GO BP terms from the DEU tissue-specific gene lists are unique to the tissue.

Our investigation shows that while there is some convergence in their GO ancestor term classification and in non-significant KEGG pathways, the majority of significantly enriched GO terms are completely separate. Therefore, not only are the genes tissue-specific but the affected processes are as well. When we do see the same pathways hit in the two tissues we mostly see different genes from these pathways affected in the muscle or the CNS. There are various pathways that are known to have participants on either side of the synapse. The converging phenotype observed when *Smn* is knocked down in either muscle or neuronal tissue along with this tissue-specificity suggests that SMA phenotype is made up of a mix of various defects; various genes, processes and pathways are impacted on either side of the NJM that build to the disease phenotype we see in humans.

In our muscle assay we were able to tune an allelic series comparable to the human variation of SMA severity types. Many previous *Drosophila* models use fly lines that create a very strong loss of *Smn* function, close to a null. Our lab's collaborators

have defined a number of fly disease models that can mimic the *Smn* dose-dependent severity of the human SMA disease. Using this model, we found a subset of the differentially expressed genes that appear to be affected by *Smn* in a dose-dependent manner in muscle: the expression level of *Smn* shows a correlated relationship to the expression level of these genes and thus we infer that they are dependent on the dose of *Smn*. This pattern of correlation can be seen in both the differential expression analysis as well as in the differential exon usage analysis, building a valuable disease model system that mimics the disease's various class phenotypes in a model organism. While it is entirely possible that there are disease relevant genes that are affected by *Smn* that do not fit the dose-dependent model pattern, this method of filtering provides strong initial candidates for study. We still see tissue-specificity when comparing these lists to our significant CNS lists. The GO BP and KEGG pathways that are further enriched in this filtered set are largely similar to those of the full significant gene list and as such many values are still unique to muscle when compared to CNS.

Our investigation of the enriched biological processes and pathways confirmed many previous findings. We had significant overlap between our lists and the combined list of genetic modifiers. When we looked at the BP GO terms based on the genetic modifiers list, we see that the genetic modifiers classification list covers almost all classifications from both the muscle and CNS DE and most of the DEU lists, including the dose-dependent filtered muscle list. When we compare the actual BP terms, we see no overlap in the DE muscle or CNS with the genetic modifiers list. However, the DEU BP term lists do have term overlap: 25% of terms and a Jaccard similarity coefficient of 0.17 in the DEU muscle terms and 27% and a Jaccard similarity coefficient of 0.18 of

DEU CNS terms. It is reassuring to see that many of our generated classifications and some terms overlap with those of established *Smn* interactors, indicating our sequencing analysis compliments previous genetic studies. Genetic validation of these RNA-Seq findings could be a focus of future studies.

Previous investigations into *Drosophila Smn* complete loss showed developmental arrest (Garcia et al., 2013). When we looked for L1 and L2 stage specific genes we did not see a significant enrichment and we concluded that our model was not found to be in developmental arrest. This shows that the variations in expression of our significant genes between sample and control are caused by the levels of *Smn*, not by overall developmental arrest. Even without developmental arrest we saw an increase in developmental GO terms in our significant gene lists. Interestingly in our differential expression analysis we did not see an enrichment of developmental GO terms with our *Smn* loss model. However, we do see an enrichment of terms relating to development in both the GSEA and DEU enriched BP GO terms.

We found many GO terms related to the functions of SMN outlined in Kolb and Kissel's review (Kolb & Kissel, 2011). Enrichment in metabolism related GO terms were found in every list, which could be linked to SMN's role in mitochondrial homeostasis and bioenergetics pathways. We also saw GO terms associated with endocytosis in the dose-dependent muscle list (both the full list and tissue-specific list), DEU muscle (full list, dose-dependent, and tissue-specific lists), and CNS significant lists. Endocytosis is essential for neuronal signaling including in the neuromuscular junction and has been seen to be impacted by SMN in previous research. Autophagy is another role of SMN that is corroborated here in our analysis, appearing in the DEU CNS

GO terms both in full and tissue-specific lists as well as in the DE muscle dose-dependent tissue-specific list. We also see GO terms related to cytoskeleton organization and biogenesis in all DEU lists and in the GSEA analysis.

Previous efforts from our lab and collaborators have found connections between *Smn* and the FGF/Ras/MAPK and BMP/SMAD pathways. Both viability and gross NMJ phenotypes can be rescued by elevating activity in these pathways (Chang et al., 2008; Sen et al., 2011). While BMP and FGF pathways were not among the overall significantly enriched pathways extracted by the KEGG analysis of our data, we did find significant changes in gene expression for individual genes that we queried from these KEGG and Flybase pathways seen in Table 2. Moreover, when we examined the more complete signaling pathways tagged by different terms, such as Ras/MAPK (downstream of FGF) or TGF-beta (of which BMP is a sub-family branch), there were additional results, seen in the table below (Table 2). We also found additional results in the wnt pathway, a well-known trans-synaptic pathway (Packard et al., 2002). While the wnt pathway did not yield results in our previous genetic screens we do see a number of our DE and DEU genes appear in the overall pathway flybase pathway as well as the KEGG pathway. The pathway nodes found in both the genetic screen and our RNA-Seq data do not need to match for us to conclude that *Smn* impacts the functional output of these pathways. In a genetic screen, mutations in any gene crucial to the pathway can appear as *Smn* genetic modifiers even if they do not directly interact with *Smn*. *Smn* needs only to impact one key element in the pathway to significantly alter its output, not necessarily the gene of interest in the screen. Each pathway has different genes affected in muscle and in CNS, further emphasizing the different specific functions *Smn* has in each tissue.

Table 2. Pathway *Smn* Interactors

	***Smn* genetic interactors that show DE/DEU in our RNA-Seq data**		
	BMP	**FGF**	
DE Muscle	*tkv, gbb*		
DE CNS			
DEU Muscle	*tkv, gbb, Trio*	*stumps*	
DEU CNS	*tkv, Trio*	*stumps*	
	All significant expression changes from our RNA-Seq data in the Flybase pathway		
	BMP (Flybase)	**FGF (Flybase)**	**Wnt-TCF (Flybase)**
DE Muscle	*tkv, tld, lili, Cul2, Smurf, cv-2, fuss, spin, magu, gbb, Lpin, ltl, Tao*	*pbl, Zpr1, Dsor1, rau, drk, troll, bnl, sxc*	*Wnt6, Swim, SkpA, rempA, ebd1, por, pbl, Usp47, Klp64D*
DE CNS	*spict, tld, magu, Ote*	*Ras85D, grh, troll, Shc*	*Wnt6, Nek2, Wnt2, ewg, wg, Oseg4, Notum*
DEU Muscle	*cv-2, tkv, tok, gbb, lili, Smurf, spin*	*Aop, drk, bnl, stumps, Mmp2, grh*	*aop, arm, sgg, ewg, Ssdp, spen, Sin3A, Mmp2, Swim, Hipk, Notum, pan, Oseg1*
DEU CNS	*eIF4A, tkv, Cul2, cmpy*	*rl, stumps*	*sgg, mts, otk, ewg, Rho1, Oseg1*
	All significant expression changes from our RNA-Seq data in the overall pathway		
	TGFBeta	**MAPK**	**Wnt**
DE Muscle	*S6k, E2f2, tkv, SkpA, Roc1b, Pp2A-29B, Smurf, gbb, achi, Rok*	*Traf6, fs(1)N, sinah, tkv, Dsor1, 14-3-3zeta, Rac1, Jra, ebi, Ask1, p38a, bsk, drk, wgn, p38b, RhoL*	*Pka-C2, sinah, Ssl, SkpA, Roc1b, p120ctn, dnt, Rac1, Jra, gskt, ebi, Plc21C, Wnt6, bsk, por, CycD, RhoL, Pp2B-14D*
DE CNS	*vis*	*Egfr, alph, Shc, eff, Rac2, Ras85D*	*Notum, wg, Ror, Wnt6, Rac2*
DEU Muscle	*Myc, tkv, Smurf, gbb*	*dos, raw, tkv, Mef2, Tak1, ttk, alph, slpr, aop, hep, Mtl, drk, eff*	*Myc, arm, Notum, CaMKII, Tak1, sgg, shf, pan, Plc21C, Mtl, Pp2B-14D*
DEU CNS	*S6k, rl, tkv, mts, Rho1, achi, vis*	*rl, tkv, msn, 14-3-3zeta, Krn, bsk, mts, Cka, chic*	*pk, DAAM, sgg, bsk, Rho1*

Not surprisingly for developmental signaling pathways, such as BMP and FGF, our DIOPT analysis shows that *Smn*-dependent genes identified in our study have a high percentage of predicted human orthologs: over 70% of our DE genes and over 80% of our DEU genes. This endorses this model as viable and human-relevant for the use of modeling SMA as a disease.

Limitations

This investigation had a few unavoidable limitations in its design based on the datasets available for analysis. First, the two datasets were sequenced on different platforms and the samples were taken at different times. Additionally, the sample sizes, while acceptable, were small, 3 biological replicates per sample (4 in the case of the muscle control). Larger sample sizes would give a better estimate of biological variation and give more accurate estimates of the mean expression levels. If we were to repeat this experiment, we would process the samples from all three genotypes at the same time in the same way, with a greater number of biological replicates.

Further Research

Further research could be done that would build upon what we have accomplished here. One recommended direction is genetic validation of some of the differentially expressed genes. The SMA phenotypic gradient model shown here in *Drosophila* will be useful in further research into the dose-dependent aspects of *Smn* function related to SMA. Another avenue of inquiry would be to obtain human SMA patient samples and use RNA-Seq to find the differentially expressed genes in a human model. This could be compared to the *Drosophila* orthologs we found here in the DIOPT analysis. In this way

one could find a conserved set of differentially expressed genes that are important to the SMA disease phenotype. Going further, differentially expressed genes from similar tissue-specific samples from SMA human patients or other animal SMA models compared to our tissue-specific results could find conserved tissue-specific functions and pathways for SMN. Far down the line after building upon the greater understanding of SMN's fundamental functions, new therapeutics that restore the most impactful pathways in specific tissues could go a long way towards healing the disease, even if not every function of SMN is rescued. For example, improving affected pathways in muscle could work in tandem with current CNS therapies to improve patient outcomes.

References

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

Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, *22*(10), 2008–2017. https://doi.org/10.1101/gr.133744.111

Andrews, S. (2010). *FASTQC. A quality control tool for high throughput sequence data*.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Carlson, M. (2019). *org.Dm.eg.db: Genome wide annotation for Fly* (R package version 3.8.2). https://doi.org/10.18129/B9.bioc.org.Dm.eg.db

Chang, H. C. H., Dimlich, D. N., Yokokura, T., Mukherjee, A., Kankel, M. W., Sen, A., Sridhar, V., Fulga, T. A., Hart, A. C., Van Vactor, D., & Artavanis-Tsakonas, S. (2008). Modeling spinal muscular atrophy in Drosophila. *PLoS ONE*, *3*(9), 1–18. https://doi.org/10.1371/journal.pone.0003209

Chaytow, H., Huang, Y. T., Gillingwater, T. H., & Faller, K. M. E. (2018). The role of survival motor neuron protein (SMN) in protein homeostasis. *Cellular and Molecular Life Sciences*, *75*(21), 3877–3894. https://doi.org/10.1007/s00018-018-2849-1

Dimitriadi, M., Derdowski, A., Kalloo, G., Maginnis, M. S., O'Hern, P., Bliska, B., Sorkaç, A., Nguyen, K. C. Q., Cook, S. J., Poulogiannis, G., Atwood, W. J., Hall, D. H., & Hart, A. C. (2016). Decreased function of survival motor neuron protein impairs endocytic pathways. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(30), E4377–E4386. https://doi.org/10.1073/pnas.1600015113

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Doktor, T. K., Hua, Y., Andersen, H. S., Brøner, S., Liu, Y. H., Wieckowska, A., Dembic, M., Bruun, G. H., Krainer, A. R., & Andresen, B. S. (2017). RNA-sequencing of a mouse-model of spinal muscular atrophy reveals tissue-wide changes in splicing of U12-dependent introns. *Nucleic Acids Research*, *45*(1), 395–416. https://doi.org/10.1093/nar/gkw731

Duffy, J. B. (2002). GAL4 system in Drosophila: A fly geneticist's Swiss army knife. *Genesis*, *34*(1–2), 1–15. https://doi.org/10.1002/gene.10150

FDA. (2019). *FDA approves innovative gene therapy to treat pediatric patients with spinal muscular atrophy, a rare disease and leading genetic cause of infant mortality*. FDA News Release. https://www.fda.gov/news-events/press-announcements/fda-approves-innovative-gene-therapy-treat-pediatric-patients-spinal-muscular-atrophy-rare-disease

Garcia, E. L., Lu, Z., Meers, M. P., Praveen, K., & Matera, A. G. (2013). Developmental arrest of Drosophila survival motor neuron (Smn) mutants accounts for differences in expression of minor intron-containing genes. *Rna*, *19*(11), 1510–1516. https://doi.org/10.1261/rna.038919.113

GARD. (2018). *Spinal muscular atrophy*. Genetic and Rare Diseases Information Center. https://rarediseases.info.nih.gov/diseases/7674/spinal-muscular-atrophy

Groen, E. J. N., Talbot, K., & Gillingwater, T. H. (2018). Advances in therapy for spinal muscular atrophy: promises and challenges. *Nature Reviews Neurology*, *14*(4), 214–224. https://doi.org/10.1038/nrneurol.2018.4

Hu, Y., Flockhart, I., Vinayagam, A., Bergwitz, C., Berger, B., Perrimon, N., & Mohr, S. E. (2011). An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics*, *12*.

Hu, Z.-L., Bao, J., & Reecy, J. (2008). CateGOrizer: A Web-Based Program to Batch Analyze Gene Ontology Classification Categories. *Online Journal of Bioinformatics*, *9*(January 2017), 108–112. http://users.comcen.com.au/~journals/geneontologyabs2008.htm

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30. https://doi.org/10.1093/nar/28.1.27

Kim, J.-K., Jha, N. N., Feng, Z., Faleiro, M. R., Chiriboga, C. A., Wei-Lapierre, L., Dirksen, R. T., Ko, C.-P., & Monani, U. R. (2020). Muscle-specific SMN reduction reveals motor neuron-independent disease in spinal muscular atrophy models. *The Journal of Clinical Investigation*, *130*(3), 1271–1287. https://doi.org/10.1172/JCI131989

Kolb, S. J., & Kissel, J. T. (2011). Spinal Muscular Atrophy: A Timely Review. *Archives of Neurology*, *68*(8), 979–984. https://doi.org/10.1001/archneurol.2011.74

Laird, A. S., Mackovski, N., Rinkwitz, S., Becker, T. S., & Giacomotto, J. (2016). Tissue-specific models of spinal muscular atrophy confirm a critical role of SMN in motor neurons from embryonic to adult stages. *Human Molecular Genetics*, *25*(9), 1728–1738. https://doi.org/10.1093/hmg/ddw044

Lefebvre, S., Burlet, P., Liu, Q., Bertrandy, S., Clermont, O., Munnich, A., Dreyfuss, G., & Melki, J. (1997). Correlation between severity and SMN protein level in spinal muscular atrophy. *Nature Genetics*, *16*(3), 265–269. https://doi.org/10.1038/ng0797-265

Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Lotti, F., Imlach, W. L., Saieva, L., Beck, E. S., Hao, L. T., Li, D. K., Jiao, W., Mentis, G. Z., Beattie, C. E., McCabe, B. D., & Pellizzoni, L. (2012). An SMN-dependent U12 splicing event essential for motor circuit function. *Cell*, *151*(2), 440–454. https://doi.org/10.1016/j.cell.2012.09.012

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

Matera, A. G., & Wang, Z. (2014). A day in the life of the spliceosome. *Nature Reviews. Molecular Cell Biology*, *15*(2), 108–121. https://doi.org/10.1038/nrm3742

NIH. (2012). *Background on Comparative Genomic Analysis*. US National Human Genome Research Institute.

Packard, M., Koo, E. S., Gorczyca, M., Sharpe, J., Cumberledge, S., & Budnik, V. (2002). The Drosophila Wnt, wingless, provides an essential signal for pre- and postsynaptic differentiation. *Cell*, *111*(3), 319–330. https://doi.org/10.1016/S0092-8674(02)01047-4

Pagliardini, S., Giavazzi, A., Setola, V., Lizier, C., Di Luca, M., DeBiasi, S., & Battaglia, G. (2000). Subcellular localization and axonal transport of the survival motor neuron (SMN) protein in the developing rat spinal cord. *Human Molecular Genetics*, *9*(1), 47–56. https://doi.org/10.1093/hmg/9.1.47

Schrank, B., Götz, R., Gunnersen, J. M., Ure, J. M., Toyka, K. V, Smith, A. G., & Sendtner, M. (1997). Inactivation of the survival motor neuron gene, a candidate gene for human spinal muscular atrophy, leads to massive cell death in early mouse embryos. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(18), 9920–9925. https://doi.org/10.1073/pnas.94.18.9920

Sen, A., Dimlich, D. N., Guruharsha, K. G., Kankel, M. W., Hori, K., Yokokura, T., Brachat, S., Richardson, D., Loureiro, J., Sivasankaran, R., Curtis, D., Davidow, L. S., Rubin, L. L., Hart, A. C., Van Vactor, D., & Artavanis-Tsakonas, S. (2013).

Genetic circuitry of Survival motor neuron, the gene underlying spinal muscular atrophy. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(26), 2371–2380. https://doi.org/10.1073/pnas.1301738110

Sen, A., Yokokura, T., Kankel, M. W., Dimlich, D. N., Manent, J., Sanyal, S., & Artavanis-Tsakonas, S. (2011). Modeling spinal muscular atrophy in Drosophila links Smn to FGF signaling. *The Journal of Cell Biology*, *192*(3), 481–495. https://doi.org/10.1083/jcb.201004016

Shen, L. (2016). *GeneOverlap: An R package to test and visualize gene overlaps*. https://doi.org/10.18129/B9.bioc.GeneOverlap

Shorrock, H. K., Gillingwater, T. H., & Groen, E. J. N. (2018). Overview of Current Drugs and Molecules in Development for Spinal Muscular Atrophy Therapy. *Drugs*, *78*(3), 293–305. https://doi.org/10.1007/s40265-018-0868-8

Singh, R. N., Ottesen, E. W., & Singh, N. N. (2020). The First Orally Deliverable Small Molecule for the Treatment of Spinal Muscular Atrophy. *Neuroscience Insights*, *15*, 2633105520973985. https://doi.org/10.1177/2633105520973985

Spring, A. M., Raimer, A. C., Hamilton, C. D., Schillinger, M. J., & Matera, A. G. (2019). Comprehensive Modeling of Spinal Muscular Atrophy in Drosophila melanogaster. *Frontiers in Molecular Neuroscience*, *12*, 113. https://doi.org/10.3389/fnmol.2019.00113

Sugarman, E. A., Nagan, N., Zhu, H., Akmaev, V. R., Zhou, Z., Rohlfs, E. M., Flynn, K., Hendrickson, B. C., Scholl, T., Sirko-Osadsa, D. A., & Allitto, B. A. (2012). Pan-ethnic carrier screening and prenatal diagnosis for spinal muscular atrophy: Clinical laboratory analysis of >72 400 specimens. *European Journal of Human Genetics*, *20*(1), 27–32. https://doi.org/10.1038/ejhg.2011.134

Tolwinski, N. S. (2017). Introduction: Drosophila-A model system for developmental biology. *Journal of Developmental Biology*, *5*(3), 10–11. https://doi.org/10.3390/jdb5030009

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118