2023

Genomic comparison of DBA/2J and C57Bl/6J strains of Mus musculus and best practice of genome alignment for bioinformatics analyses

Dustin J. Zeliff
Virginia Commonwealth University

Genomic comparison of DBA/2J and C57Bl/6J strains of *Mus musculus* and best practice of genome alignment for bioinformatics analyses

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics at Virginia Commonwealth University

by

Dustin Zeliff
Bachelor of Science, Virginia Commonwealth University, 2020

Michael F. Miles, M.D., Ph.D.
Professor, Department of Pharmacology and Toxicology

Virginia Commonwealth University
Richmond, Virginia
July 21, 2023

Contents

18

Acknowledgements

I have many people I want to thank, but none more so than Dr. Michael Miles, who has been my advisor and mentor throughout this project. You have been an excellent mentor and example for me, and I am incredibly grateful to have been able to work with you these last few years. You've helped me understand what it means to be a bioinformatician and to rise to the expectations of the field. I will be using the skills and habits you've taught me for the rest of my career, and I am truly grateful for that.

I would also like to acknowledge Dr. Mikhail Dozmorov, for all of his help with the data preparation end of my project. I ran into several stumbling blocks while using new data, which are detailed in this paper, but you were always more than happy to help me work through them, and you taught me so much about working with large, confusing data sets and files. I am honored to have you be a part of my project and a member of my committee.

I would like to acknowledge the other members of my committee as well, Dr. Paul Fawcett and Dr. Karolina Aberg. Dr. Fawcett, you taught me more about how to learn and grow as a programmer than any other professor I've had. I definitely started off in over my head in your class, but you helped me work through it and improve while still helping me learn. After speaking to you after class and in office hours many times, I knew I wanted to have you on my committee for my thesis. I wanted people I knew would be tough on me and push me to succeed. Thank you for that, and for teaching me nearly everything I know about python. Dr. Aberg, I didn't know you personally before Dr. Johnson recommended you for my committee, but I am so grateful that she did. Your feedback on my proposal and ideas have helped me improve my both

my scientific work and my presentation skills, and I am truly thankful you have been a part of my committee.

Dr. Allison Johnson, thank you for all of your advice and assistance through these last few years. You made me feel welcomed and supported throughout my master's program, and I cannot thank you enough for everything you've done on my behalf. Whether it was reaching out to me to talk about my application, helping me select a mentor for my thesis, or helping me navigate the morass of course scheduling, I've always felt that you were there to support me in whatever it was I needed help with. For that I am truly grateful.

I'd like to thank the Miles Laboratory, firstly Emma Gnatowski. You've helped me understand everything about my project, from the code to the results. You've been a great mentor and colleague, and this project is just as much yours as it is mine. Thank you for everything, I don't think I would have succeeded without your help.

Dr. Maren Smith, thank you for helping me with code issues and for talking to me about bioinformatics and the world beyond the lab. You helped me gain some perspective and understanding about bioinformatics and what the world is looking for out of us.

Walker Rodgers, Sam Gotlieb, Zachary Tatom, Doug Bledsoe, and Analise Hassan, thank you all for helping me with anything I needed. You were always willing to help with problems I was having, or just to talk about our projects and help me understand the research you were doing. Not to mention all of your help with my presentation, you guys definitely helped me prepare to defend my proposal and thesis!

And lastly, my parents, Dr. James Zeliff and Dr. Tamara Zeliff. Thank you for all of your support. It’s certainly been a long, winding journey to get here, but you’ve been there every step of the way. I am truly grateful to you for everything you’ve done.

Contributions

I am honored to have received aid from several colleagues during the course of my thesis. Any contributions not listed in the text are found below.

Data Preparation

Dr. Mikhail Dozmorov assisted in the generation of count data from the D2 mouse annotation file provided by Dr. Keane. This proved to be a very tricky task, but we managed to accomplish it together.

Tables

Chapter 2

Chapter 4

Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Tools and Programs

Bioinformatics Tool	Link
STAR Aligner	https://github.com/alexdobin/STAR
MultiQC	https://multiqc.info/
Deseq2	https://bioconductor.org/packages/release/bioc/html/Deseq2.html
ToppFun	https://toppgene.cchmc.org/enrichment.jsp
Revigo	http://revigo.irb.hr/
DEXSeq	https://bioconductor.org/packages/release/bioc/html/DEXSeq.html
RegTools	https://regtools.readthedocs.io/en/latest/
rMats	https://RNA-Seq-mats.sourceforge.net/
List Comparison	https://rnact.crg.eu/compare
List Comparison	https://comparetwolists.com/
BioVenn	https://www.biovenn.nl/
Pheatmap	https://www.rdocumentation.org/packages/pheatmap/versions/1.0.12/topics/pheatmap
Ggplot2	https://ggplot2.tidyverse.org/

Abstract

Alcohol use disorder is known to have significant genetic components that contribute to an individual's susceptibility to the disease. Mouse models are commonly used to study the mechanisms underlying alcohol use disorder, with C57BL/6J (B6) and DBA/2J (D2) being two of the more prominently used inbred strains. Research in the Miles Laboratory has used these two strains, and genetic panels of mice derived from them, to identify potential genes associated with variance in ethanol-related behaviors using quantitative trait loci (QTL) analysis. For example, Ninein (Nin) was identified as a potential candidate gene for the anxiolytic effects of ethanol, discovered because it resides in the confidence interval for a QTL and shows mRNA expression differences between B6 and D2 mice. This differential expression was identified using counts of RNA-Seq reads that have been aligned to a reference genome, specifically the B6 reference genome. Due to the known genetic differences between the two strains, it is possible that the D2 samples could benefit from being aligned to a D2 genome instead of the B6. This would lead to better results overall due to improved read alignment and identification of novel splicing events that might be seen in D2 mice. To test this hypothesis, a dataset consisting of deep (150 million reads) sequencing of RNA from nucleus accumbens of both B6 and D2 mice was used for multiple bioinformatics analyses (differential expression, gene ontology, semantic similarity, differential exon utilization, splice site location, and alternative splicing) with both B6 aligned D2 counts and D2 aligned D2 counts. End results of each analysis were then compared for significant differences in outcomes. The results of this analysis show that when aligning D2 samples to the D2 genome a majority of differentially expressed genes and differentially utilized exons are retained from the B6 aligned analysis while many new genes and exons are identified that are unique to the D2 aligned analysis.

Chapter 1: Introduction and Background

Introduction

Alcohol Use Disorder

Alcohol Use Disorder (AUD) describes the spectrum of problematic alcohol consumption that affects over 29 million people in the United States (SAMHSA, 2021). AUD includes increased alcohol consumption over time and binge alcohol consumption, though it encompasses any kind of problematic alcohol use. All forms of AUD relate to the inability to regulate or stop alcohol use despite external pressures such as negative social, health, or occupational consequences. AUD also leads to multiple alcohol-related end-organ diseases, affecting virtually every organ system, including such prevalent problems as fetal alcohol syndrome and alcoholic liver disease. It is estimated that 140,000 people die from AUD every year (SAMHSA, 2021), and alcohol use costs the United States $249 billion annually, with $28 billion of that coming only from healthcare costs (Sacks et al., 2015). Furthermore, less than 10% of people suffering from AUD in the past year received any form of treatment for it (Han et al., 2021), thus highlighting the need for improved understanding of the disorder so as to develop new therapeutic agents. The study of AUD has revealed it to have a genetic component, with twin studies being used to estimate that 50% of the risk of developing AUD is due to genetic factors (Kranzler et al., 2019). Single nucleotide polymorphism (SNP) based estimates are closer to 12%, and it is believed that AUD's genetic heritability is a result of many genes having small effects (Kranzler et al., 2019). Very few variants that cause changes to protein structure and function have been identified, and variants that regulate gene expression have been put forward as a potential mechanism that affects these complex traits. Alcohol produces long lasting cellular changes in the brain, and it is these changes that can eventually lead to AUD (Egervari et al., 2019) However, studying gene expression in the human brain is difficult due to the complexity of the human brain, and the scarcity of human brain tissue. Because of these limitations, model organisms are used instead.

Mus musculus as a model organism and its use in AUD research

Model organisms have been extensively studied and have well-characterized genetic, physiological, and behavioral traits. One of the more commonly used model organisms for AUD research is the house mouse, *Mus musculus*. Mice make appealing model organisms due to their genetic, physiological, anatomical, and reproductive similarities to humans, as well as more practical reasons such as the relative ease of caring for them in a laboratory environment and the vast wealth of tools and resources available for working with mice (García-García, 2020).

Inbred mice are defined as being the product of at least 20 generations of brother X sister mating, with all individuals being derived from a single breeding pair. Inbred mice have several traits that make them ideal for research purposes. They are isogenic, and homozygous at each genetic locus. They have very unified phenotypes due to this stability. Due to this, inbred strains have very well documented traits, allowing for specific strains of mice to be selected for specific types of research (Blake et al., 2021). In AUD research, the C57BL/6J and DBA/2J inbred strains of mice are commonly used. This is due to several of the known traits that differ between the two strains being ideal for alcohol research, including their high variance in baseline ethanol consumption, with C57 consuming much more alcohol voluntarily than D2, and .

C57BL/6J, more commonly referred to as C57 or B6, are the most widely used inbred strain. They are often used as a background strain for behavioral genetic studies in alcohol research due to their facile self-administration of high amounts of alcohol. B6 mice were the DNA source for the first high quality draft sequence of the mouse genome and thus were the first strain to have their genome sequenced (Waterston et al., 2002). Due to this, their genome is one of the most well studied, and is widely used as the standard alignment sequence for genomic analyses. As Figure 1.1 shows, they are particularly useful for alcohol research as they voluntarily consume large quantities of alcohol (Lê et al., 1994).

DBA/2J, or D2, are the oldest of all inbred strains. They are used as a contrast to B6 mice in alcohol research, as they do not voluntarily consume large amounts of alcohol (Lê et al., 1994). Because they are so often used, the behavioral and genetic differences between the two strains are well documented, especially when it comes to alcohol research. He et al. (1997)

performed an examination of these differing traits and their genetic components.

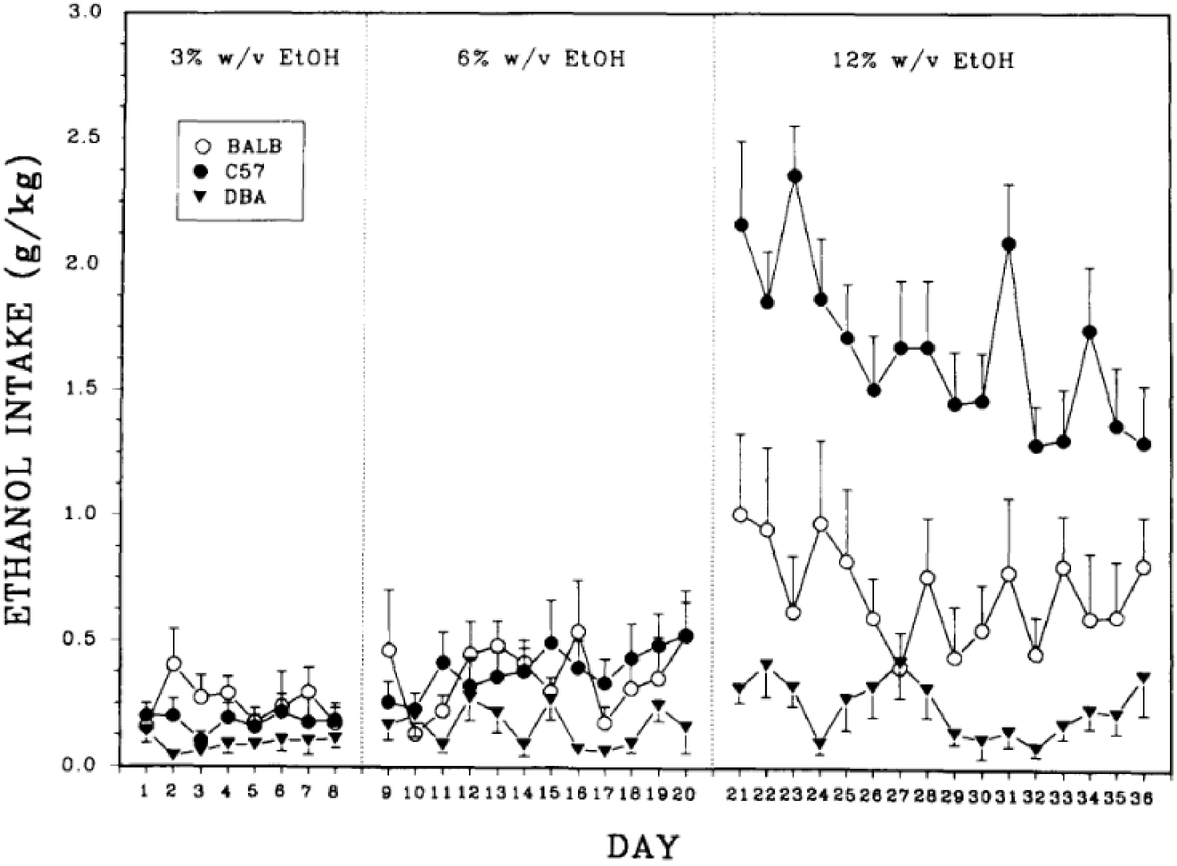

Figure 1.1. Alcohol intake (g/kg) by C57BL/6, BALB/c (another inbred strain of mice), and DBA/2 mice during the l-h daily access to alcohol solution. The concentration of alcohol solution was 3% w/v for the first 8 days, 607o for the next 12 days, and 12070 for the remaining 16 days. N = 17-18 mice per strain. Vertical lines indicate positive or negative halves of the SEs. Figure and description from Lê et al. (1994).

RNA-Sequencing

RNA-Sequencing is a technique used to measure gene expression in cells or tissues. The output of RNA-Sequences is a series of reads that represent the expression levels of individual genes. These reads are often short and fragmented, which makes it difficult to know where they came from in the genome. In order to utilize these reads, they must be aligned to a reference genome. A reference genome must be a high quality, well-annotated representation of the genome. Through the alignment process, the locations that the reads originated in the genome can be identified. Once the reads are aligned, the total reads that overlap between the sample and the reference genome are counted, which quantifies the expression level of each gene in the sample (Martin & Wang, 2011).

Previous research and inspiration for this study

Miles laboratory studies have included extensive genome-level expression studies (Kerns et al., 2005a) (Agarwalla et al., 2020) and behavioral genetic analyses across the B6, D2 and recombinant (BXD) mice. Behavioral genetic analysis across the BXD recombinant inbred panel was used to identify genetic quantitative trait loci (QTL) modulating the anxiety-reducing actions of ethanol (Putman et al., 2016). Microarray gene expression across the BXD mice was further used to identify possible candidate genes for the QTL (Wolen & Miles, 2012). This analysis has recently shown that the gene Ninein (Nin), located within a highly significant behavioral quantitative trait locus (QTL) contributing to the anxiolytic-like properties of ethanol, was differentially expressed between B6 and D2 mice and that there was possible differential exon utilization for Nin expression between the two strains (Putman et al., 2016). Ninein is a gene that codes for a microtubule binding protein that is important in axonal development and is known to interact with Gsk3β. (Srivatsa et al., 2015). It was suggested to be a possible candidate gene for alcohol's anxiolytic effects (Putman et al., 2016).

Statement of significance

The B6 genome has been used as the reference genome for the majority of mouse studies, and virtually all RNA-Seq analysis, as it is the highest quality and best annotated genome available. However, there are known genetic differences between the B6 and D2 genomes. In addition to the research done in the Miles laboratory, initial sequencing efforts of the D2 genome have identified over five million single-nucleotide polymorphism and insertion/deletion differences between B6 and D2 mice (Doran et al., 2016). These genetic differences may lead to lower quality alignment when sequencing data from D2 are aligned to the B6 reference genome, compared to when data from B6 are aligned to the same reference. Which in turn may lead to biased results for downstream analyses. In particular, this difference may complicate studies on differential exon utilization.

Roadmap and Hypothesis for this study

The genetic variation and differential expression shown between the two strains provides a basis for the hypothesis that aligning D2 mice to their own genome will show a significant difference in outcomes when compared with aligning D2 mice to the B6 genome.

Using a recent deep-sequencing RNA-Seq dataset obtained in the Miles laboratory for B6 and D2 mice, I analyzedseveral kinds of bioinformatics studies between B6 and D2 reads aligned to the B6 reference genome versus results using D2 reads aligned to a recently derived D2 reference genome. The analyses to be performed are differential expression, gene ontology, differential exon utilization, and differences in splicing. If there are significant differences in results using D2 aligned D2 samples compared to the results using the B6 aligned D2 samples, then this will allow for better analyses by aligning to the D2 genome instead of the B6. If there are significant effects caused by aligning to the D2 genome, that also opens more avenues of research into other strains of mice.

Specific Aims

This project has two specific aims, both furthering the overall goal of comparing analyses run with D2 aligned D2 samples to those run with B6 aligned D2 samples. First, there will be a comparison of differential gene expression and gene ontology between the two strains of mice and an analysis of how aligning the D2 mice to their own genome changes those results. This will further the understanding of the effect alignment has on the results of gene expression analyses. The gene ontology will be used to compare the results of the two differential expression analyses at a functional level, and a semantic similarity analysis will continue that goal to further compare the semantic groupings of gene ontology categories.

Second, I will be comparing differential exon utilization and alternative splicing between the two strains, analyzing how aligning the D2 mice to their own genome changes those results. The comparison will be using DexSeq (Differential EXon and Transcript analysis for RNA-Seq) to compare exon utilization and alternative splicing, respectively. DexSeq is a computational method for detecting differential exon usage in RNA-Seq data, and is an extension of differential expression analysis, instead identifying differences in the usage of individual exons or groups of exons between samples.

This will show the abundance of alternative splicing events, and will focus on specific genes to showcase the differences in alternative splicing on a gene level caused by aligning D2 mice to the D2 genome. A gene ontology of the genes with differentially expressed exons will also be performed. This ontology will determine the most specific functions of each gene with differentially utilized exons, to further understand the differences caused by the change in alignment.

This research can lead to future studies to determine the impact of differential exon utilization on the proteome. Analyses can be conducted to determine how many different protein-coding sequence elements are derived when aligning to the D2 genome vs the B6 genome. Other future goals include a deeper look into the alternative splicing and changes in splicing events between the B6 and D2 aligned analyses, and a breakdown of the differentially utilized exons by size and other factors to determine if there is a pattern in the exons missed or picked up by the two analyses.

Chapter 2: Sample and Data Preparation

Introduction

Sample preparation

The samples for this study were prepared before this study began, following the procedures outlined in the methods section. The RNA-Seq data that came out of that work was the inspiration for this project, as the deep sequencing allowed for a robust analysis of the differences between the B6 and D2 strains.

Alignment and Count Generation using D2 genome

D2 alignments have been attempted before in the Miles Laboratory, but the relatively low quality of the prior existing D2 sequence data and genome annotations has made them less efficient than the consensus B6 annotations for RNA-Seq alignments. In some cases, the lower quality of the D2 annotations made certain analyses impossible to perform with samples and counts aligned to them. In this study a new high quality D2 genome sequence and annotation was provided by Dr. Thomas Keane from the Sanger Center. This sequence and annotation are of a high enough quality to allow for alignments at a similar level to those using the B6 genome. The annotation initially had Ensembl IDs corresponding to the D2 genome, whereas the B6 aligned counts had IDs corresponding the B6 genome. This issue was rectified using the annotation file, which contained gene names mapped to the D2 IDs. These gene names were mined from the file using a series of python scripts (Appendix 2 – ID Conversion Scripts), then converted to B6 Ensembl IDs, and mapped to the D2 IDs. The D2 IDs in the newly generated count files were then replaced with their corresponding B6 Ensembl IDs, in order to perform Deseq2 and DexSeq analyses.

Methods and Materials

Sample Preparation

In initial studies conducted on ethanol regulation of Ninein gene expression by Jessica Jurmain during the course of her M.S. thesis work in the Miles Laboratory (2020), eight-week old male C57Bl/6J and DBA/2J mice were obtained from Jackson Laboratories (Bar Harbor, ME). The mice were housed in cages on ventilated racks with Teklad Sani-Chip bedding (currently Envigo, Cumberland, VA) and cotton nesting material. Four mice were housed in each cage. A 12-hour light dark cycle was maintained at all times and the mice were fed ad-libitum with Teklad LM-485 7012 standard rodent chow and tap water. Two weeks after the mice had arrived, they were given 0.9% saline, 1.8 g/kg or 4 g/kg ethanol via intraperitoneal injection and then euthanized 4 hours later by cervical dislocation and decapitation. This was done to obtain brain tissue from the nucleus accumbens for dissection and subsequent molecular studies. This tissue was the source of RNA used for the RNA-Seq studies that form the basis of this work. All procedures were approved by the Virginia Commonwealth University Institutional Animal Care and Use Committee in accordance with National Institute of Health guidelines.

Immediately following decapitation, the entire brain was removed and microdissected as described by Kerns et al. (2005). Briefly, the whole brain tissue was chilled on ice for 1 minute in 1x phosphate buffer then dissected by sectioning and micropunch to isolate tissue from 7 regions of the brain, including the nucleus accumbens. The tissue samples were then placed in individual tubes, flash frozen using liquid nitrogen, and stored at -80 degrees Celsius until RNA extraction.

RNA was extracted from the nucleus accumbens tissue using homogenization in STAT-60 (Tel-test, Inc., Friendswood, TX, USA) and purified with a Qiagen RNeasy Mini Kit (Qiagen, Redwood City, CA, USA). A ThermoFisher Nanodrop 2000 Spectrometer was used to assess RNA concentration by measuring the UV-Vi's absorbance at 260 nm. The sample quality was assessed using Agilent Technologies Agilent RNA 6000 Nano Kit. Samples with RNA quality indicator (RQI) values less than 7.0 were not used. The control samples (saline-treated) from B6 and D2 mice (n=5/strain) were then prepared for RNA-Sequencing at the VCU genomics core facility by Emma Gnatowski in in the Miles Laboratory and provide the resource for the analysis performed in this study.

D2 Annotation File Preparation

The annotation file provided by Dr. Keane was initially in GFF3 format. While this format will work with STAR aligner (Dobin et al., 2013) for the generation of counts, the SAMSORT (Danecek et al., 2021) and DexSeq (Anders et al., 2012) applications both require GTF files. In order to convert the GFF3 file to a GTF file, a docker environment was created and AGAT (Another Gtf/Gff Analysis Toolkit) (Dainat et al., 2020) was used to convert the GFF3 file to a GTF file. This worked for SAMSORT, but for DexSeq an additional step was required. GTF files do not usually contain parent relationships for the genes and transcripts contained within, but GFF3 files do. This causes the "Parent" attribute to conflict with the "gene_ID" and "transcript_ID" attributes. Removing the "Parent" attributes leaves the file with the same attributes as a normal GTF file, which was needed to prepare the DexSeq counts.

DESeq2 Count Generation

The following steps were performed using the VCU Group high performance computing cluster. The FASTA file for the D2 genome taken from the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/view/GCA_921998315.2) was modified so that the headers matched the chromosome names of the GFF3 file. Then STAR aligner was used to generate an index file for the count generation process using the FASTA file and GFF3 file (Appendix 2: submit02a_STAR_index.sh). Next, the samples were aligned to the D2 genome using STAR aligner, generating BAM files (Appendix 2: submit02b_STAR.sh). The indexed BAM files were then sorted with SAMSORT (Appendix 2: SortScript.sh). These sorted BAM files would be used directly in the DexSeq count preparation, explained further in chapter 4. Feature counts would then be generated for the Deseq2 analysis using the converted GTF file and the sorted, indexed BAM files, explained further in chapter 3.

Results

D2 Annotation File Preparation

The initial GFF3 (Appendix 1: DBA_2J_v3.2.gff3) file was successfully converted to a GTF file using AGAT. The resulting file is DBA_2J_v3.2_3_14_23.gtf (Appendix 1). It was then successfully prepared for DexSeq analysis, with the resulting file being DBA_2J_v3.2_3_14_23_filtered.gtf (Appendix 1).

Deseq2 Count Generation

The headers of the FASTA file were successfully changed to match the GFF3 and GTF file chromosome names (Appendix 1: GCA_9219983152_FASTA_Converted_DZ_3_23_23.fasta). STAR aligner successfully generated the index files (Appendix 1 – Index Files) followed by the BAM files (Appendix 1 – BAM Files). Finally, SAMSORT successfully sorted the BAM files (Appendix 1 – Sorted Files). MultiQC was run on the indexed BAM files to determine the percentage and number of uniquely mapped reads, the STAR alignment scores, and gene counts of each sample (Figures 2.1, 2.2, 2.3). At this stage MulitQC was also performed on the generated feature counts used in the DESeq2 analysis. This was then compared to the MultiQC results of the B6 aligned B6 and B6 aligned D2 samples (Table 2.2) using a T-test.

These same steps were performed on the B6 mouse samples that were aligned to the B6 genome, successfully generating BAM files (Appendix 1 – BAM Files). RNA-Seq samples were aligned to release 108 of the B6 reference genome using STAR aligner (Dobin et al., 2013) on the VCU group server. The B6 reference genome (https://ftp.ensembl.org/pub/release-110/fasta/mus_musculus/dna/Mus_musculus.GRCm39.dna.primary_assembly.fa.gz) and annotation (https://ftp.ensembl.org/pub/release-110/gtf/mus_musculus/Mus_musculus.GRCm39.110.gtf.gz) were taken from Ensembl (European Microbiology Laboratory - European Bioinformatics Institute, Cambrige, UK). The D2 samples were then aligned to a D2 reference genome (Assembly GCA_921998315.2) taken from the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/view/GCA_921998315.2) and annotation (DBA_2J_v3.2.gff3) provided by Dr. Thomas Keane. The resulting BAM files were checked for quality using MultiQC, and compared to the MultiQC results of the B6 alignments (Tables 1 & 2.)

Table 2.1: D2 aligned D2 samples MultiQC results showing uniquely mapped reads, both alignment percentage and millions of reads (M) and the assignments of feature counts in percentage assigned and millions of reads assigned.

Sample Name	% Assigned	M Assigned	% Aligned	M Aligned
D11N_S1_001	72.80%	111	93.60%	139.1
D13N_S8_001	71.60%	103	92.80%	127.1
D22N_S7_001	72.40%	108.4	94.10%	139.4
D32N_S10_001	72.10%	107.6	93.90%	133.7
D34N_S6_001	72.60%	99.2	94.70%	141.8
Average	72.3%	105.84	93.82%	136.22

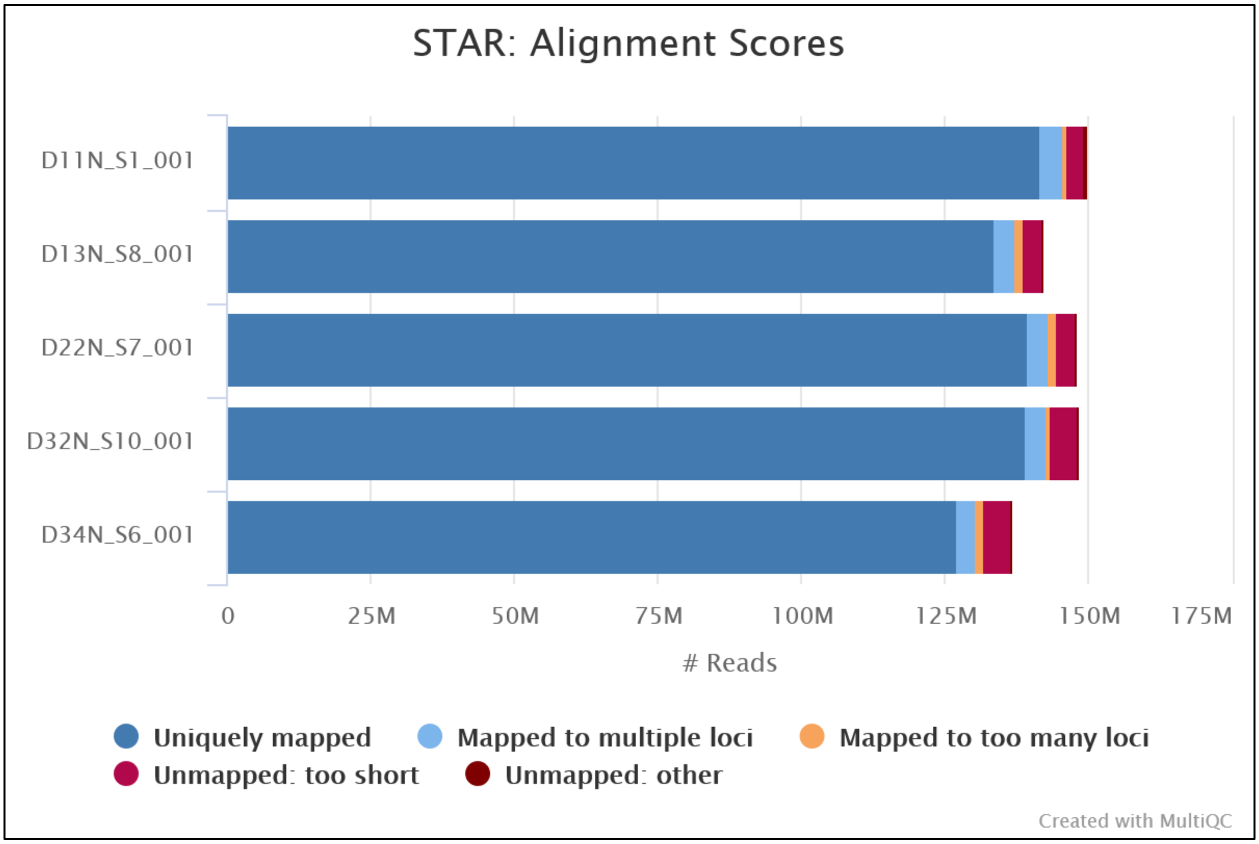

Figure 2.1: STAR alignment scores of D2 aligned D2 samples, in millions of reads.

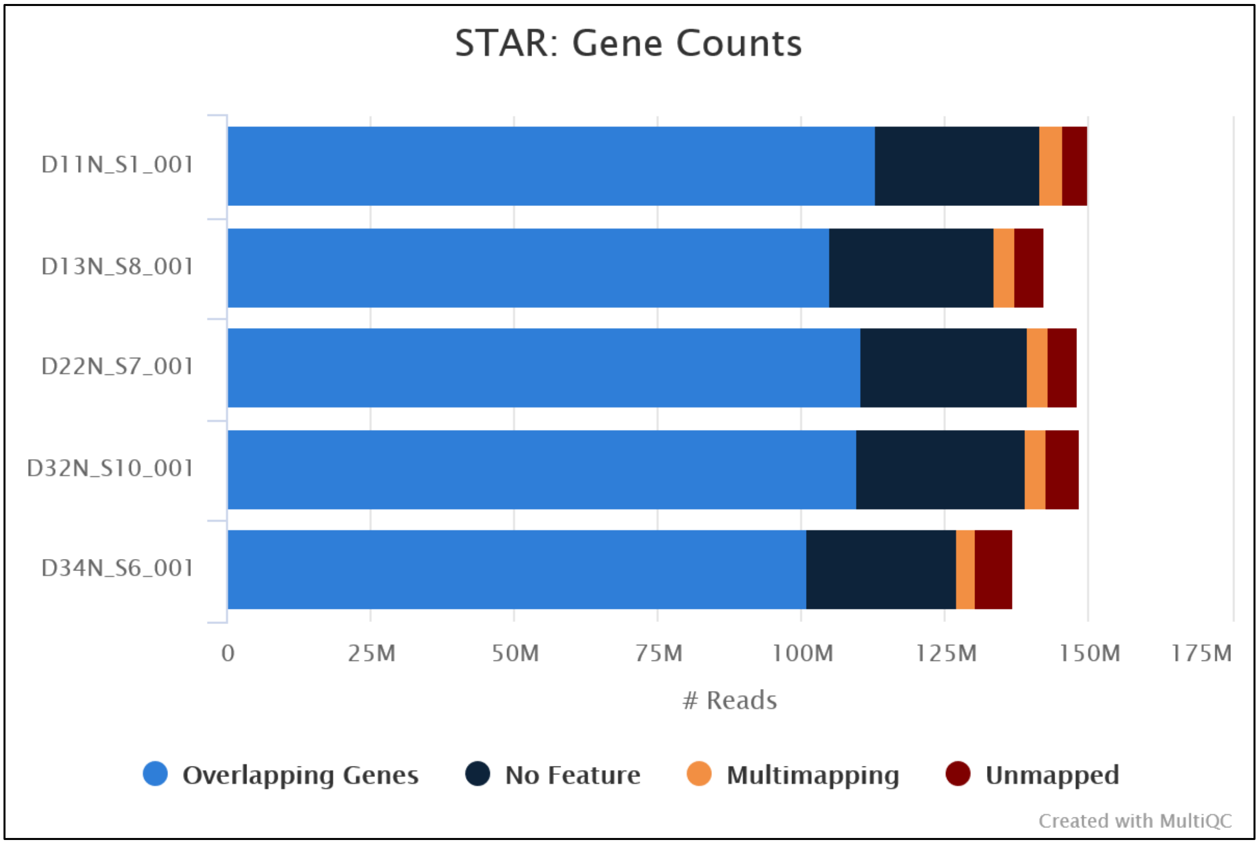

Figure 2.2: STAR gene counts of D2 aligned D2 samples, in millions of reads.

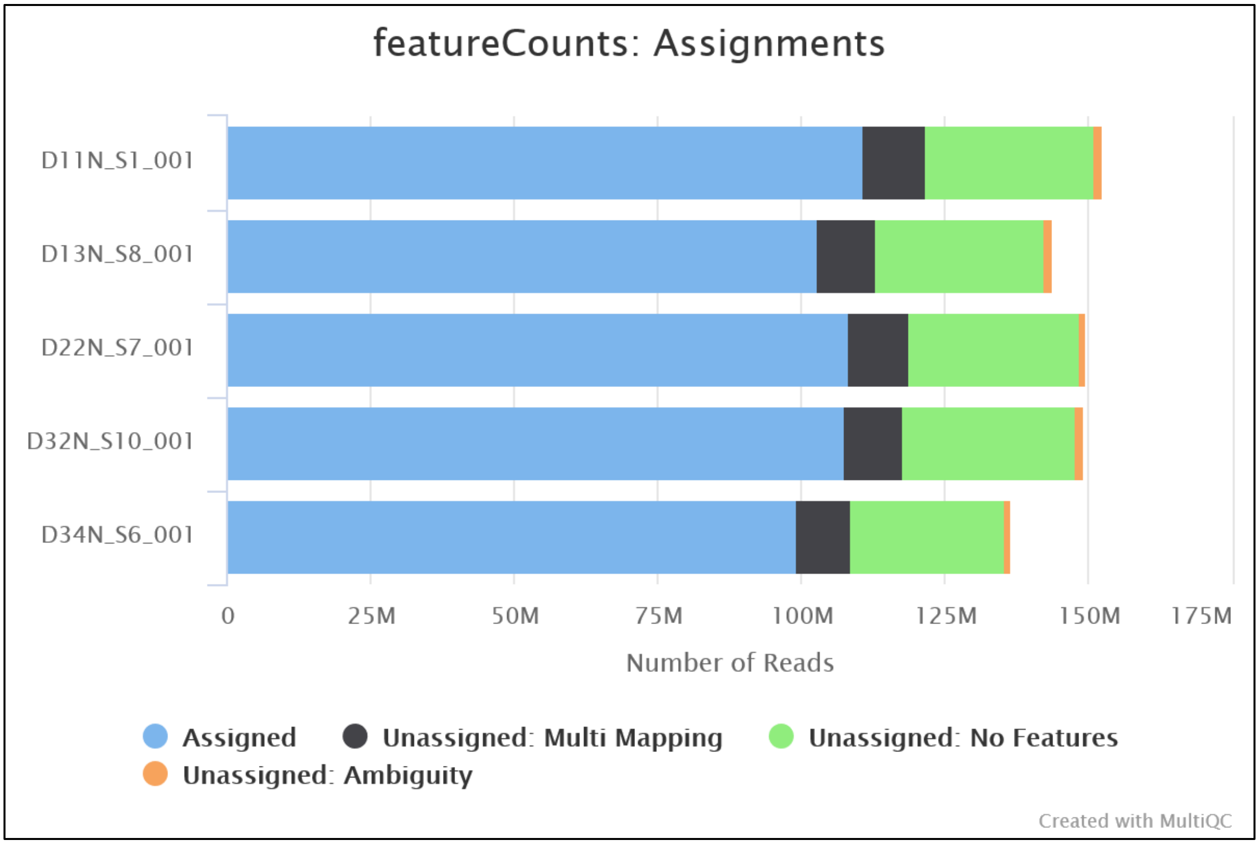

Figure 2.3: MultiQC of feature counts of D2 aligned D2 samples, showing the number of assigned features and the number of unassigned features with the reason they were not assigned.

Table 2.2: B6 aligned B6 and D2 samples showing uniquely mapped reads, both alignment percentage and millions of reads (M) and the assignments of feature counts in percentage assigned and millions of reads assigned.

Sample Name	% Assigned	M Assigned	% Aligned	M Aligned
B14N_S9	74.40%	111.5	92.90%	133.6
B21N_S5	74.90%	104	91.60%	124.1
B24N_S3	75.40%	97	90.20%	114.4
B31N_S4	74.60%	101.4	90.90%	120.9
B32N_S2	75.90%	108.2	92.00%	127.5
D11N_S1	75.00%	116.8	92.50%	138.7
D13N_S8	73.60%	108.7	92.40%	131.6
D22N_S7	74.40%	114.3	92.40%	136.9
D32N_S10	74.20%	113.2	91.60%	136.2
D34N_S6	74.60%	104.4	91.20%	124.9
B6 Average	75.04%	104.42	91.52%	124.1
D2 Average	74.36%	111.48	92.02%	133.66

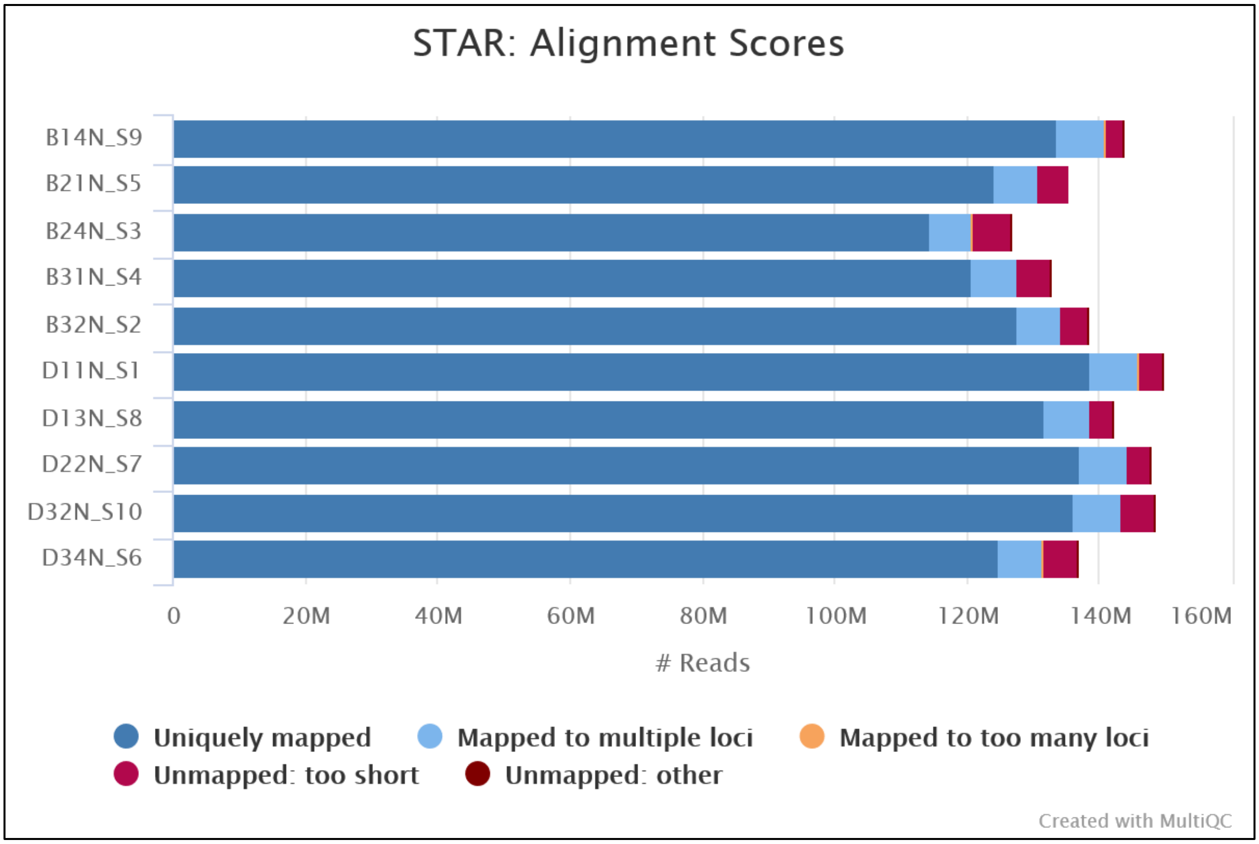

Figure 2.4. STAR alignment scores of B6 aligned B6 and D2 samples, in millions of reads.

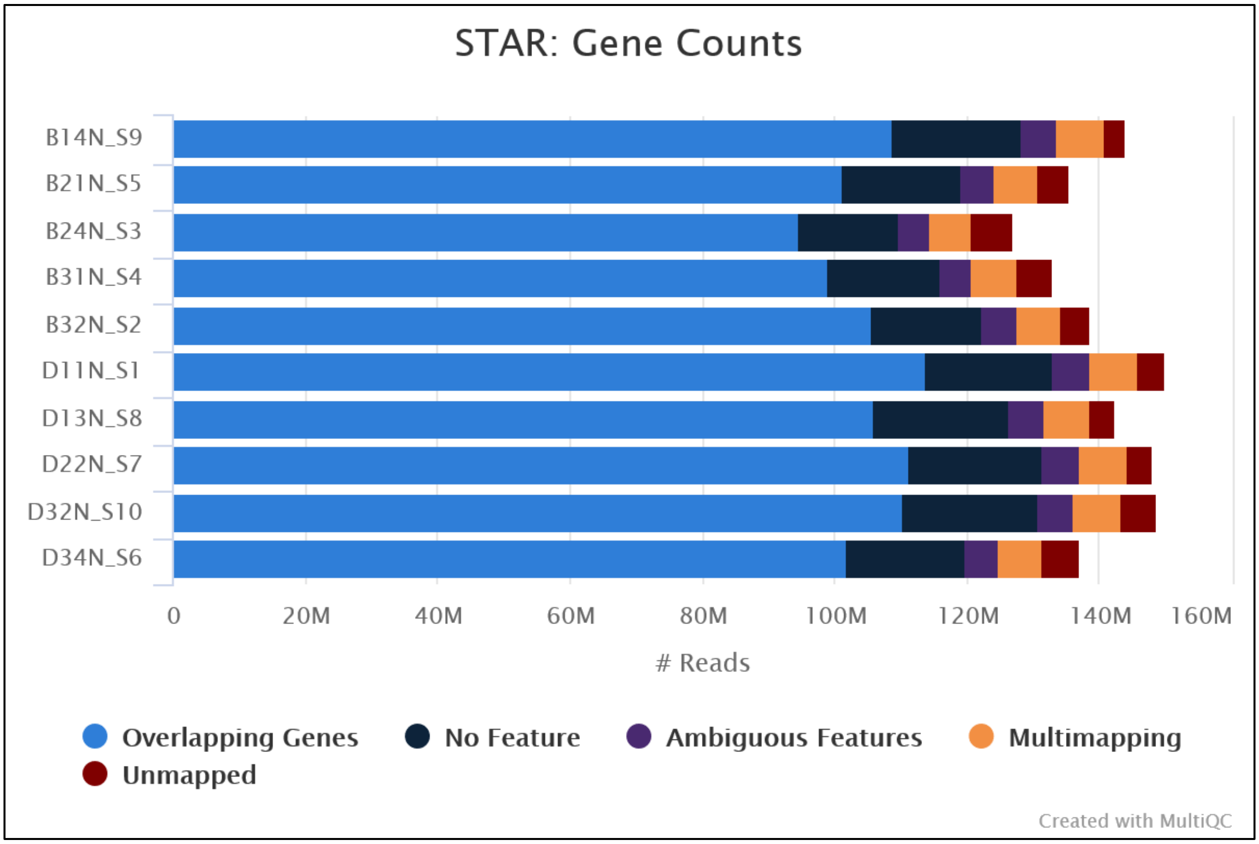

Figure 2.5. STAR gene counts of B6 aligned B6 and D2 samples, in millions of reads.

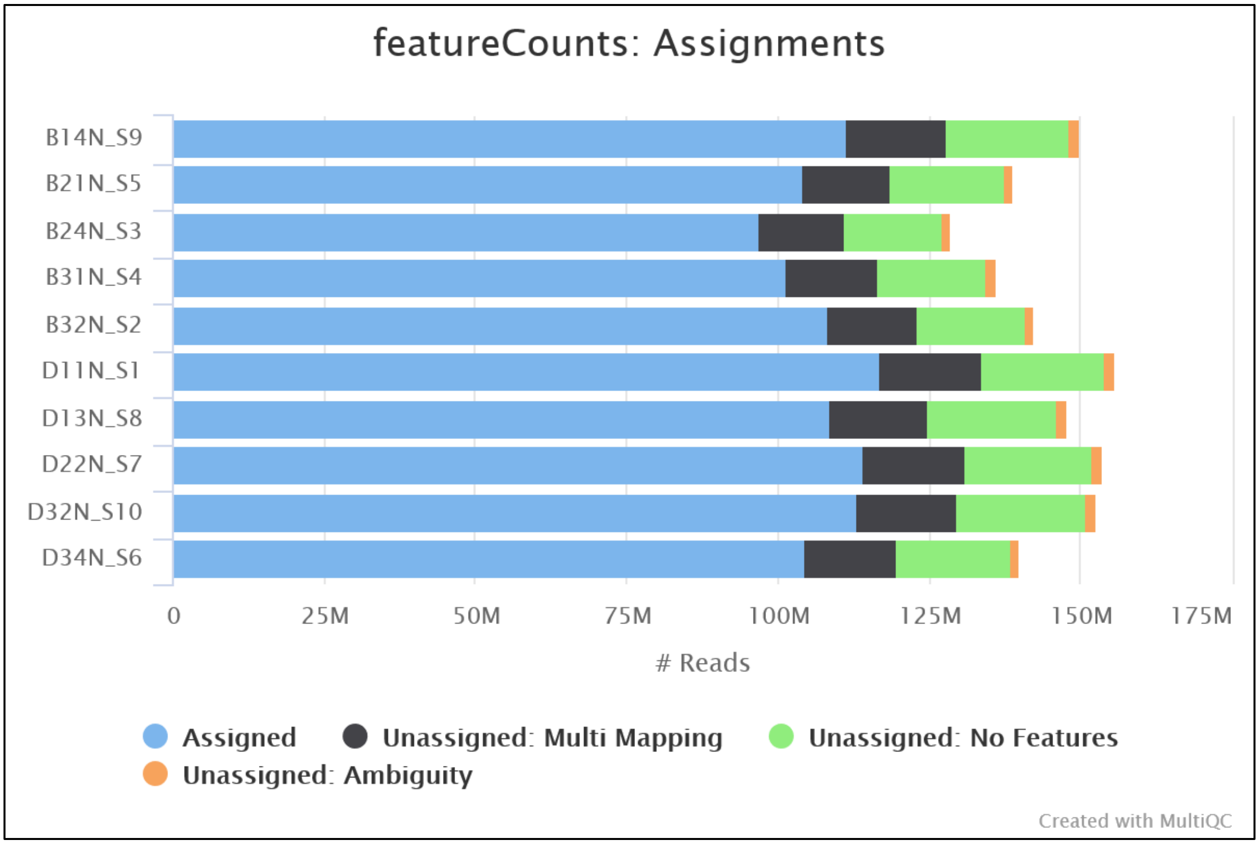

Figure 2.6. MultiQC of feature counts of B6 aligned B6 and D2 samples, showing the number of assigned features and the number of unassigned features with the reason they were not assigned.

Discussion

The D2 aligned D2 samples showed significantly less percentage and total number of uniquely mapped reads assigned with 72.3% compared to 74.36% in the B6 aligned D2 ($p < 0.0001$) and 105.84 million compared to 111.48 million total reads in the B6 aligned D2 ($p < 0.0001$). However, the D2 aligned D2 showed significantly more percentage of reads aligned with 93.82% compared to 92.02% in the B6 aligned D2 ($p = 0.0272$). The difference between the total reads aligned between D2 aligned D2 and B6 aligned D2 was not significant, with 136.22 million compared to 133.66 million total reads aligned ($p = 0.5354$). STAR aligner suggests that 80-90% alignment is acceptable, and their benchmark for experimental data is 94% aligned (Dobin et al., 2013). These results fall inside that window, and therefore the alignment percentage is acceptable. The alignment results are also higher than those used in previous differential gene expression studies that were aligning to the B6 genome (Bottomly et al., 2011), (Mortazavi et al., 2008), and with a significantly higher alignment percentage this should improve results of analyses done using these counts. The D2 alignment did produce a lower percentage of assigned reads than the B6 aligned. This could be due to several reasons, such as the complexity of the genomic regions or genetic variation between the samples and the D2 reference genome, though the most likely reason is that the D2 reference genome is less complete than the B6 reference genome. Regions that aren't well represented in the D2 reference genome would cause their associated reads to be assigned at a lower rate or not at all. However, the % assignment is still high. There is no guideline for what an acceptable assignment percentage is, however, being within 2% of the B6 aligned results is good enough to proceed. The benefits of aligning to the D2 genome, such as increasing the future analyses' ability to detect SNPs and small indels, and potential allele specific expression differences outweigh the slight decrease in assignment percentage moving forward.

Chapter 3: Differential Expression Analysis and Gene Ontology

Introduction

The process of information taken from a gene being used to create a functional product is called gene expression. This leads to the related phenotypes being shown in the resulting organism, and therefore in any kind of genetic research understanding gene expression is extremely important. Gene expression is tightly regulated (Ptashne & Gann, 1997) as any dysregulation can quickly lead to disease (Esteller, 2007). Differential gene expression is when a gene in two or more samples has a statistically significant difference in expression levels, or read counts (Anjum et al., 2016). RNA-seq data is commonly used to identify differentially expressed genes (Li & Xie, 2013) by their read counts.

Deseq2 is an R package developed by Love et al. (2014) that performs differential expression analysis on RNA-seq feature count data using a negative-binomial (Gamma-Poisson) distribution. The input data required are some form of gene identifier, Ensembl IDs were used in this study, and read counts for each sample. It goes through three steps to perform the analysis, first normalizing the data by estimating size factors, then estimating the dispersion, then running the negative binomial test. The relevant output of the analysis are p-values indicating whether a gene is significantly ($p < 0.05$) differentially expressed between the sample groups, and a log2fold change, indicating the magnitude of the differential expression (Love et al., 2014). Deseq2 is used in this analysis to compare 5 B6 aligned B6 samples and 5 B6 aligned D2 samples, then again to compare those same 5 B6 aligned B6 samples with 5 D2 aligned D2 samples, resulting in a list of significantly differentially expressed genes between the two strains.

B6 and D2 mice are known to have differential expression of genes between them and previous analyses have been run aligning to the B6 genome as it was the only available mouse genome (Bottomly et al., 2011). Being able to align the D2 mice to their own genome allows for a differential expression analysis to be performed with more accurate results, as aligning to the D2 genome will account for genetic variation (SNPs, indels) specific to that strain. In addition, reliance on a single reference genome can cause bias in downstream analyses. It can also result in the analysis missing important genetic variants if they occur in regions not present in the reference genome (Kim et al., 2019).

Gene ontology (GO) categorizes genes based on the function of their products. There are three main categories, biological processes, molecular function, and cellular components. Each category contains a hierarchy of terms, with the most specific terms at the bottom and broader terms at the top. Genes are associated with the most specific term that accurately describes their products. GO is particularly useful when comparing genes across species or, in this case, strains within a species, as it allows for a comparison of function in a set of genes (Ashburner et al., 2000). In this study, gene ontology is used to compare the functions of the significantly differentially expressed genes between B6 and D2 samples, and between the B6 and D2 aligned analyses.

Revigo is a tool that was developed by Supek et al. (2011) that is designed to take an input of gene ontology terms and their significance levels in the form of p-values and return a reduced, clustered visualization of those terms based on their semantic content. This quantifies how much the terms share a common meaning, and uses SimRel to assign a score to each based on their semantic similarity, with scores of .9 or higher indicating high similarity. This reduces the number of gene ontology terms into larger categories, making for easier visualization and comparison. SimRel is a functional similarity measure used to compare two GO terms with each other (Shlicker et al., 2006) (Figure 3.1). It is based on SimRes, Resnik's semantic similarity algorithm (Resnik, 2011) and SimLin, Lin's semantic similarity (Lin, 1998). Resnik's method focuses on the most informative common ancestor of the GO terms, and Lin's approach adds a focus on the shared information between the two terms. SimRel combines these approaches to incorporate relevance similarity (Schlicker et al., 2006).

$$sim(t_1, t_2) = \frac{2 * \log p(MIA)}{\log p(t_1) + \log p(t_2)} * (1 - p(MIA))$$

Figure 3.1. SimRel algorithm. t_1, t_2 refer to the gene ontology terms being compared, which are the most specific terms possible for each gene. p(t_1) and p(t_2) refer2 to the probability of those terms being found in the GO dataset, and p(MIA) refers to the probability of finding the common ancestors of terms t_1 and t_2 in the GO dataset. This is then weighted with 1-p(MIA) because the relevance of a term decreases with increasing probability. Equation taken from Schlicker et al., 2006

This section of the study focuses on comparing the differential gene expression between the two strains using Deseq2, then generating GO terms using ToppFun and reducing them for visualization with Revigo. This analysis will be run twice, once using B6 aligned D2 samples and once using D2 aligned D2 samples. The results of both analyses will then be compared using 2 tailed t-test to determine if there is a significant difference in the magnitude of the LFCs of filtered significantly differentially expressed genes ($p < 0.05$, FDR 0.1) and list comparison to determine the changes in differentially expressed genes identified when aligning D2 samples to the D2 genome. GO terms will then be compared using list comparison and Revigo clustering to determine if aligning the D2 samples to the D2 genome causes a significant difference in the functions of those genes' products.

Methods

Differential Gene Expression

The paired end counts generated in the previous step were run through a differential expression analysis using Deseq2 (Bioconductor) as described by Love et al. (2014). First the B6 counts that were aligned to the B6 genome were compared to D2 counts that were aligned to the B6 genome in terms of log2fold change (LFC) using Deseq2. This showed to what degree each gene was differentially expressed between the two genomes. Then, B6 counts that were aligned to the B6 genome were compared to D2 counts that were aligned to the D2 genome in terms of log2fold change. Finally, the significantly differentially expressed genes from each comparison were compared to each other using a two tailed t-test to determine if there was a significant difference in LFC between the two sets of differentially expressed genes, then using rnact.crg.eu's list comparison feature to determine the differences in which genes were differentially expressed.

Genes with median counts of less than 1 across all 10 samples were filtered out of the data. The counts were normalized using Deseq2's median of ratios method (Love et al., 2014), and pairwise correlation values were calculated for these samples. These were visualized using a hierarchal heatmap of correlation data created using pHeatmap. The pairwise correlation values for all samples were visualized using multiple scatterplots. A principal component analysis of the variance was run on the top 500 and top 10,000 genes by counts. Then the differential expression between the strains was calculated using Deseq2. The data was filtered again, taking only genes that were significant at p = 0.05 and filtered using an FDR of 0.05, again using Deseq2. These were visualized using both a volcano plot made with GGplot2 and a heatmap made with pHeatmap.

Gene Ontology and Semantic Similarity Analysis

The filtered, significantly differentially expressed genes from the previous step were used to run a gene ontology analysis using ToppFun (ToppGene, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA). The results of this gene ontology in the biological process, cellular component, and molecular function categories were then put through Revigo's semantic similarity analysis as described in Supek et al. (2011) to better visualize the groupings of genes inside those categories. Scatterplots were created using the GGPlot2 R package and treemaps were created using the treemap R package.

This was repeated for the comparison of B6 counts aligned to the B6 genome to D2 counts that were aligned to the D2 genome.

Comparison of Results

The resulting differentially expressed genes from both the analysis using B6 aligned D2 and the analysis using D2 aligned D2 were compared using simple list comparison metrics. The differentially expressed genes were compared in both number and name, with similarity being measured by how many genes were differentially expressed in both comparisons and by which genes were differentially expressed. The gene ontology results were compared to each other directly, with the total number in each category being compared as well as how similar the individual genes' functions were. This was accomplished by comparing the names directly and seeing what percentage of overlap there was between the two studies. The positive and negative sets of LFC values was determined to have significantly unequal variance ($p < 0.05$) and as such the t-tests used were Welch's t-tests, assuming unequal variance. Running a 2 tailed t-test on the significantly expressed genes from each analysis with positive LFC values, and a 2 tailed t-test on the significantly expressed genes from each analysis with negative LFC values. These were separated as the overall average of LFCs from both analyses was nearly zero, and as such would not be a good comparison. Finally, the Revigo results were compared, to see if the gene ontology results fell into similar or different broad categories.

Results

Aim 1a – Differential Gene Expression

Differential Gene Expression Between B6 Aligned B6 and B6 Aligned D2

The initial correlation data of the counts showed that the two strains were closely related, with a minimum correlation value of .992. There was also clear delineation between B6 and D2, with each sample having significantly higher correlation with samples of the same strain than samples of the other strain (Figure 3.2). The principal component analysis of the top 10,000 most abundant genes by counts showed that strains were clustered together by 80% variance (Figure 3.3). However, both Figure 3.2 and Figure 3.3 indicate that 2 B6 samples showed slight variance compared to the other B6 samples in terms of correlation and PC2 grouping. We elected to not exclude these from further analysis since their overall correlation was more similar to B6 than D2 samples, and they clustered tightly with B6 samples on hierarchical clustering and principal component analysis (PC2). A MA plot of genes with differential expression (FDR $\leq$ 0.05) was used to visualize results, showing the log10 fold-change (LFC) of all genes plotted versus the mean of normalized counts (Figure 3.4). 6,210 genes were differentially expressed (D2 vs. B6), with 3,257 having a positive log2fold change and 2,953 having a negative LFC. A positive LFC indicates that the gene showed higher expression in the D2 strain than the B6 strain. A heatmap of LFCs by genotype was generated to show the differences in LFC for each gene and each individual (Figure 3.5), with positive LFC values indicating higher expression in D2 mice. The top 20 differentially expressed genes exhibited no bias towards either positive or negative LFC (Figure 3.6), suggesting adequate normalization of the data and no systematic errors biasing the analysis.

B6 Aligned Heatmap of Correlation Data

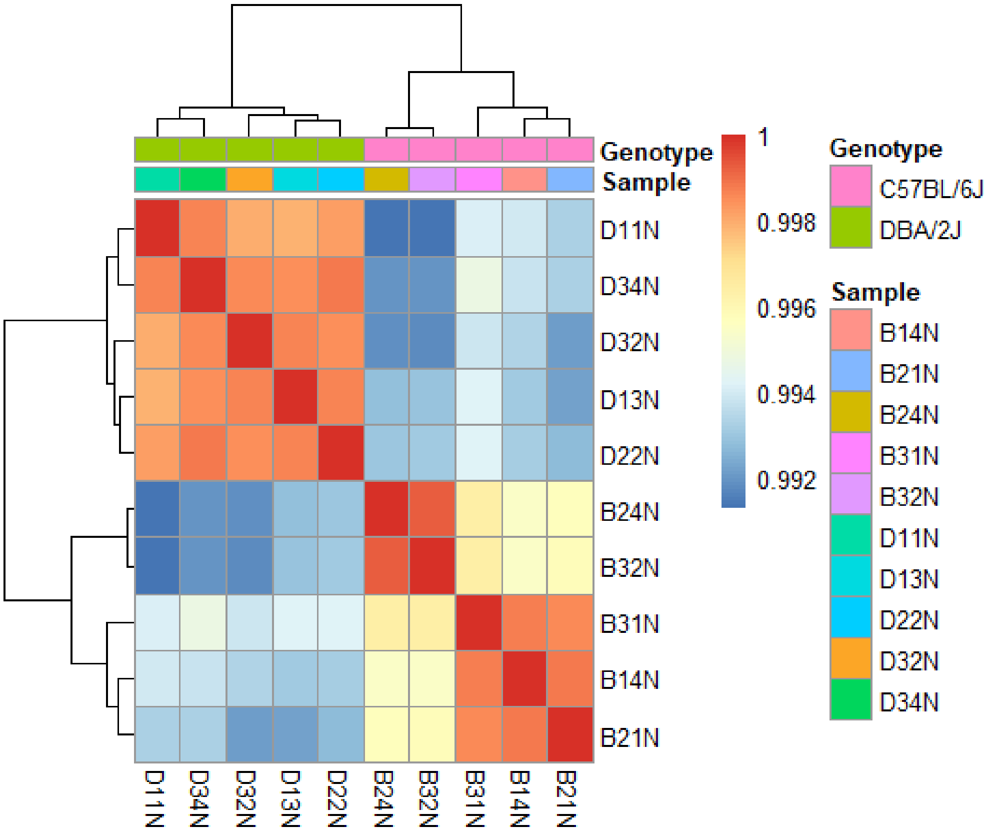

Figure 3.2. Heatmap of count correlation data of the B6 aligned analysis. B6 and D2 correlate more with themselves than with each other. There are no major outliers. The overall high levels of correlation between B6 and D2 (.992 to 1) shows clear separation between two closely related strains. Two samples, B24N and B32N, showed slightly lower correlations with the other B6 samples but still clustered tightly with the remaining B6 samples.

Principal Component Analysis of the Top 10,000 Most Abundant Genes, B6 Aligned Analysis

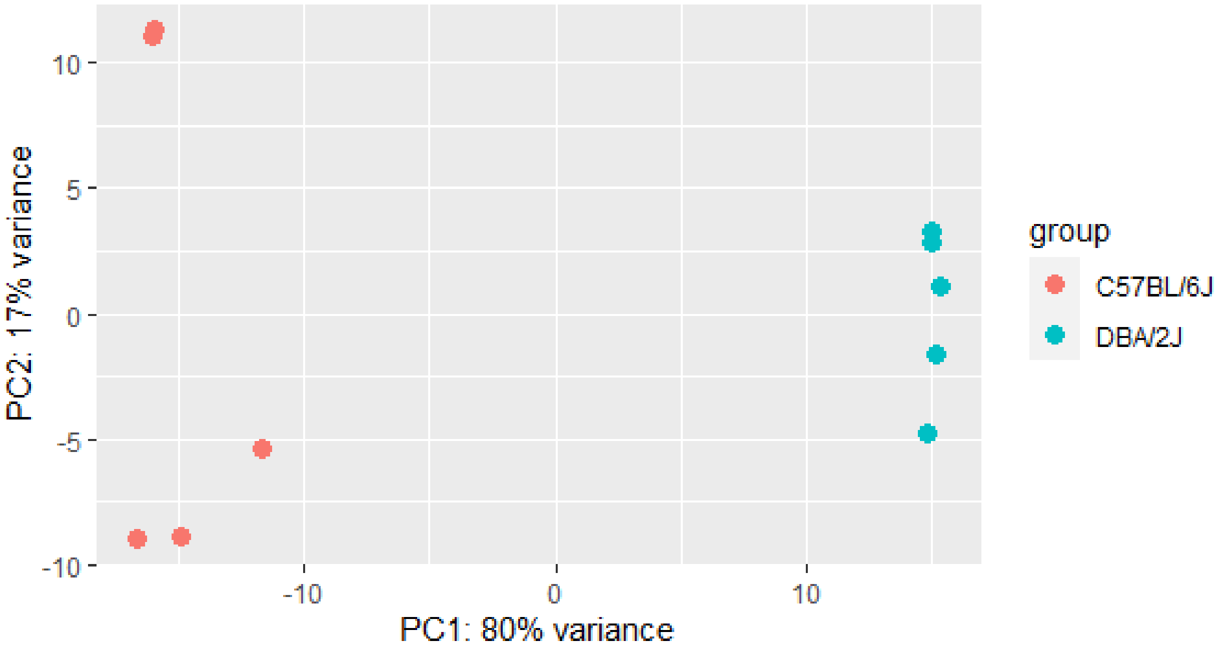

Figure 3.3. Principal component analysis of the top 10,000 most abundant genes. Substrains are clustered together along the X axis (PC1) while some variation between samples within a strain are differentiated on the Y axis (PC2).

B6 Aligned Log2Fold Change by Mean of Normalized Counts

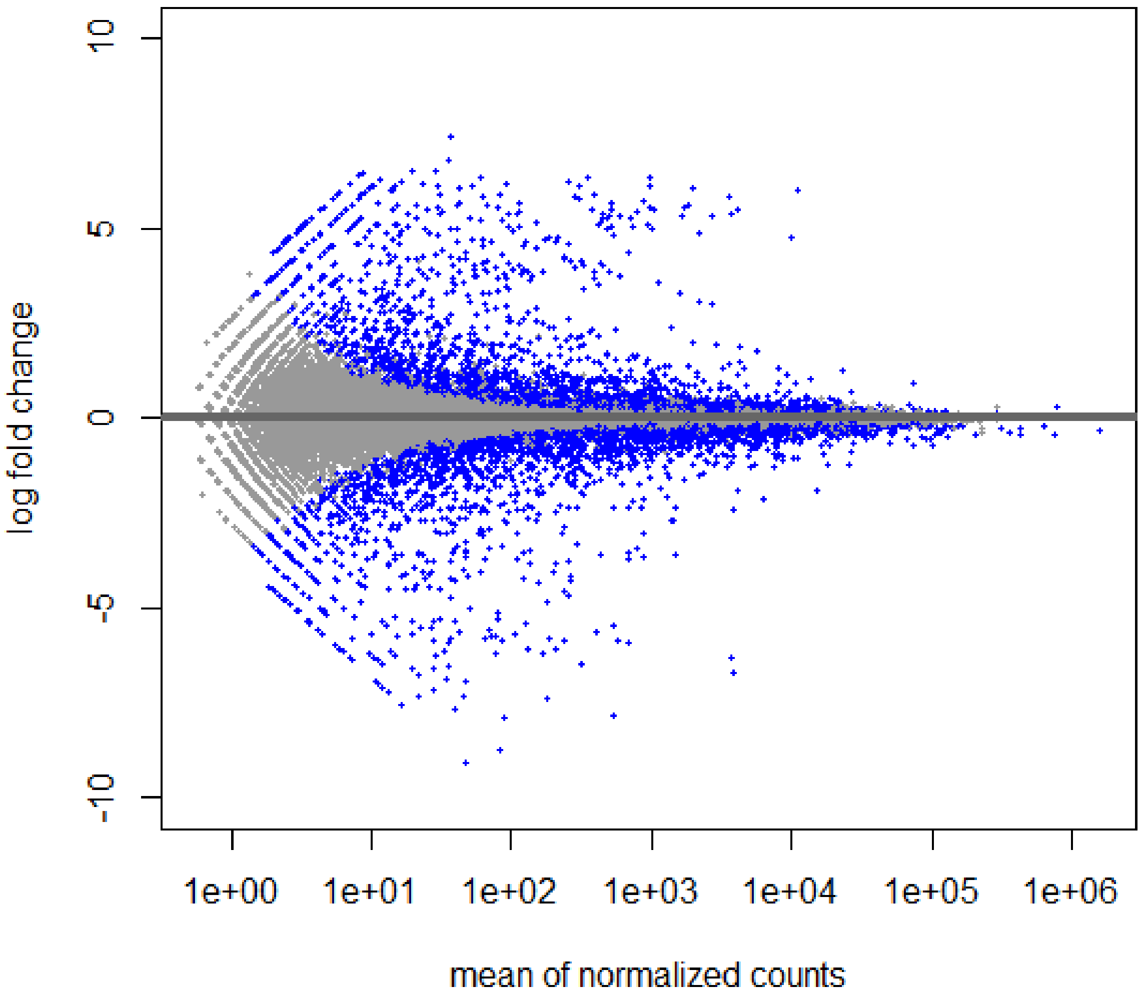

Figure 3.4. MA plot of the LFC against the mean of normalized counts for all genes. Blue indicates genes that were significantly differentially expressed between the two strains (FDR <0.05) and grey indicates genes that were not significantly differentially expressed.

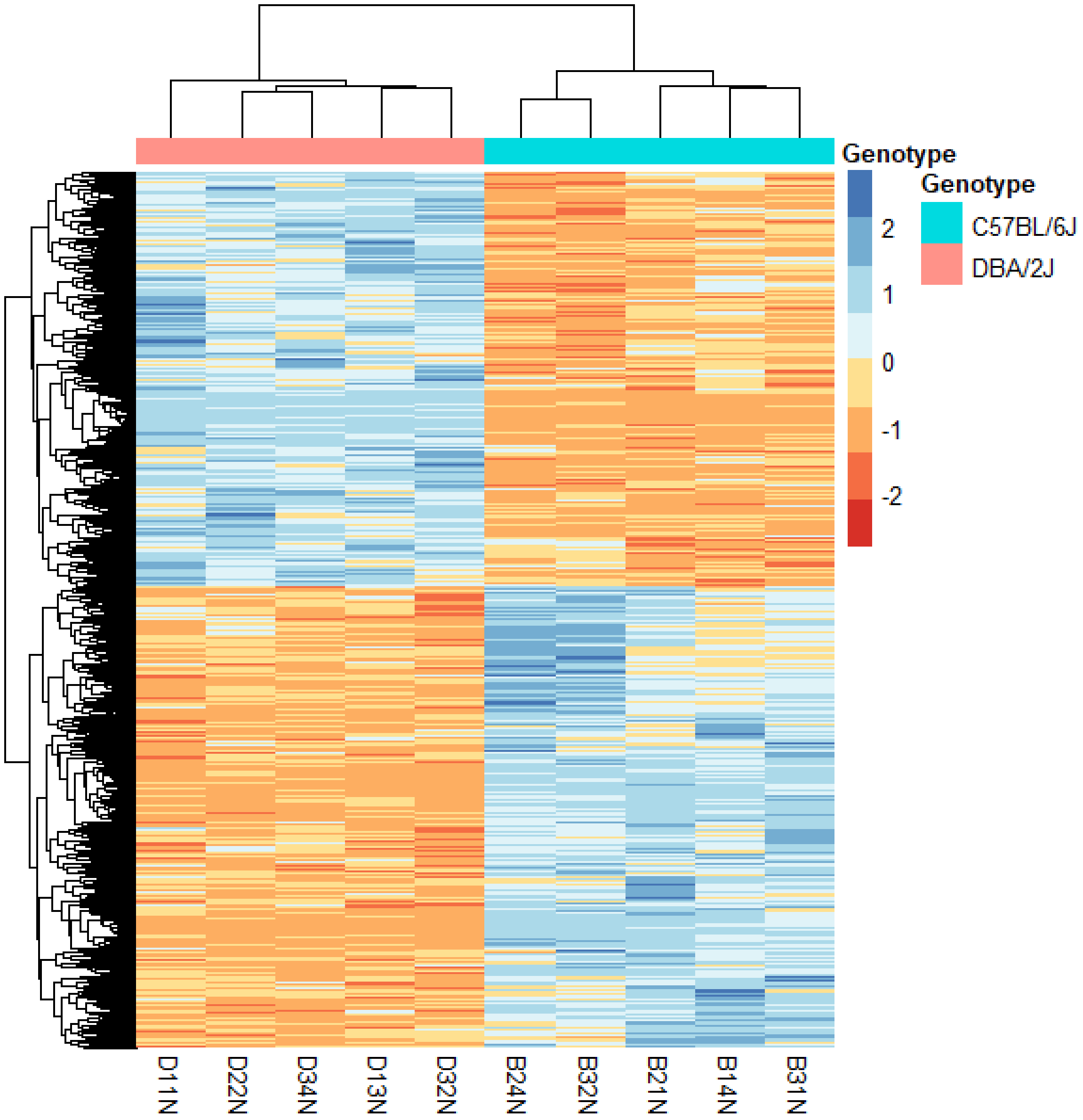

Figure 3.5. Heatmap of hierarchical cluster analysis of differentially expressed genes between B6 aligned B6 and B6 aligned D2 and the log2fold changes (LFCs) of each gene. A positive LFC (Red) indicates higher expression in D2. The 2-dimensional cluster analysis reveals robust consistency across the samples for differential expression analysis.

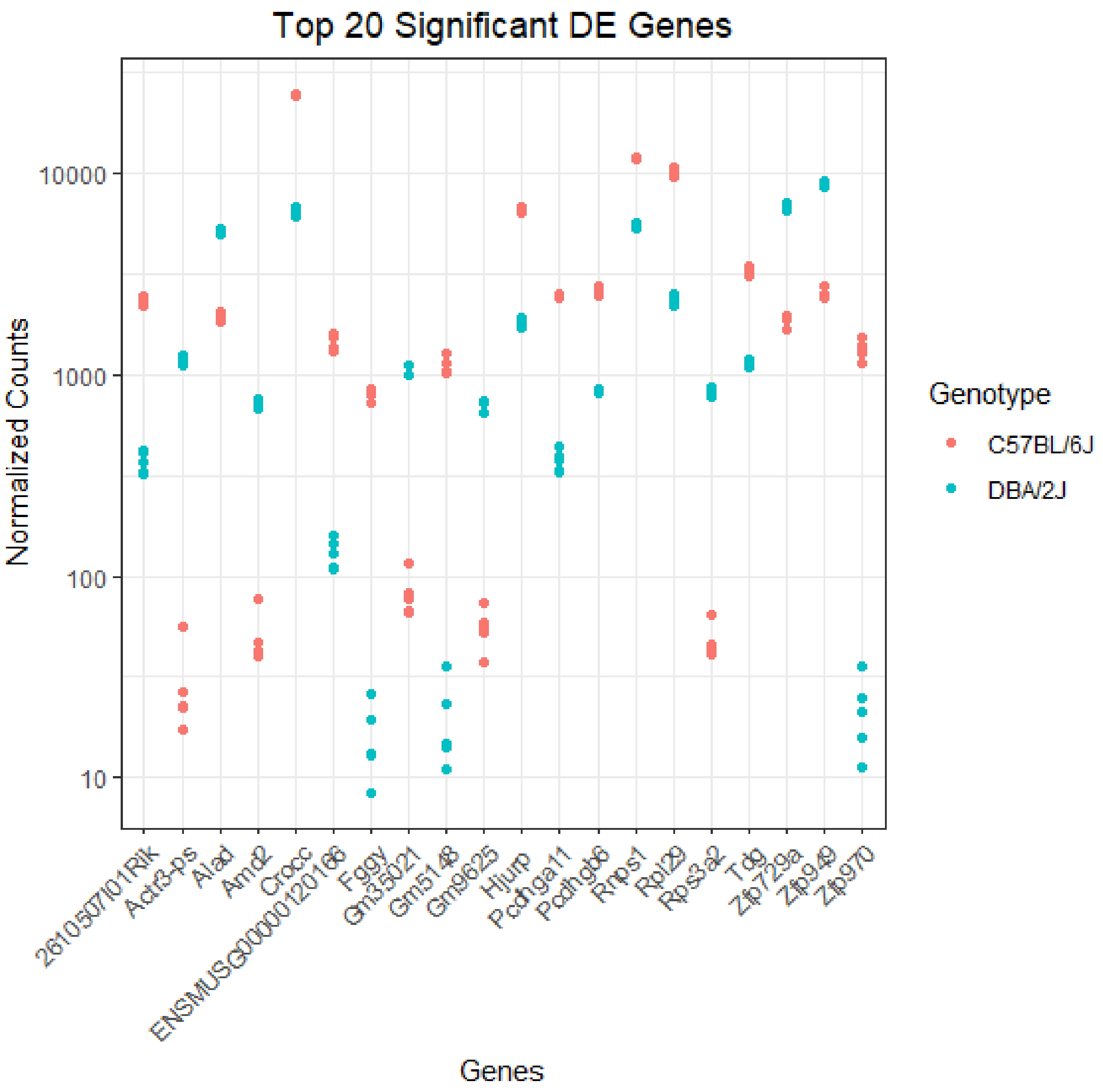

Figure 3.6. Top 20 significantly differentially expressed genes and their normalized counts in each strain. 12 genes have positive LFC values, and 8 have negative LFC values for D2 expression versus B6 expression.

Differential Gene Expression Between B6 Aligned B6 and D2 Aligned D2

Principal Component Analysis of the Top 10,000 Most Abundant Genes

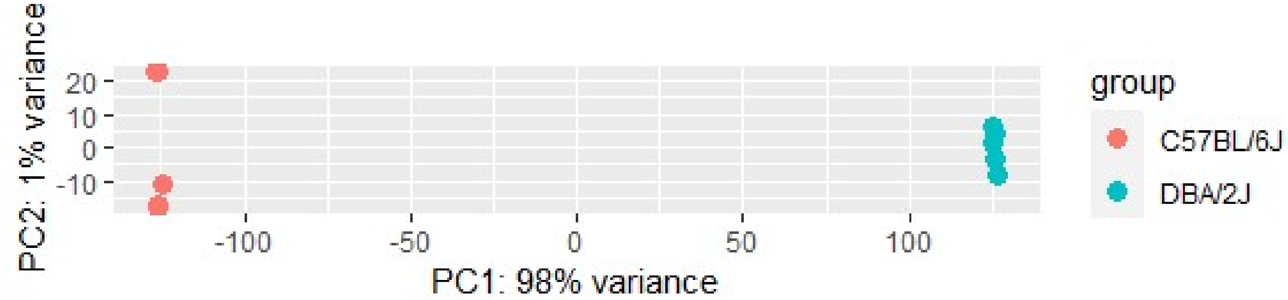

Figure 3.7. Principal component analysis of the top 10,000 most abundant genes when using the D2 aligned D2 counts. Substrains are clustered together along the X axis.

Log2Fold Change by Mean of Normalized Counts

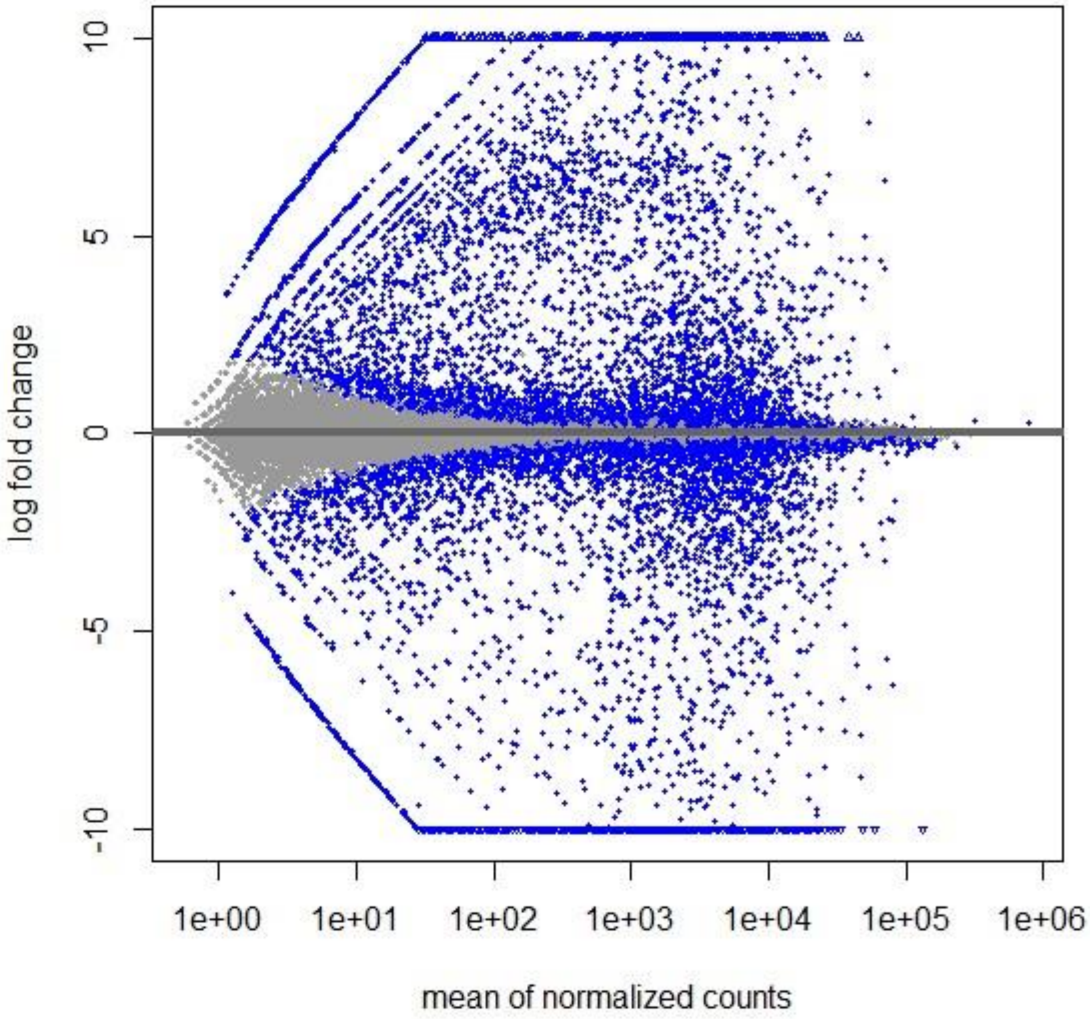

Figure 3.8. MA plot of the LFC against the mean of normalized counts for all genes. Blue indicates genes that were significantly differentially expressed between the two strains (FDR <0.05) and grey indicates genes that were not significantly differentially expressed.

Heatmap of LFC of Results

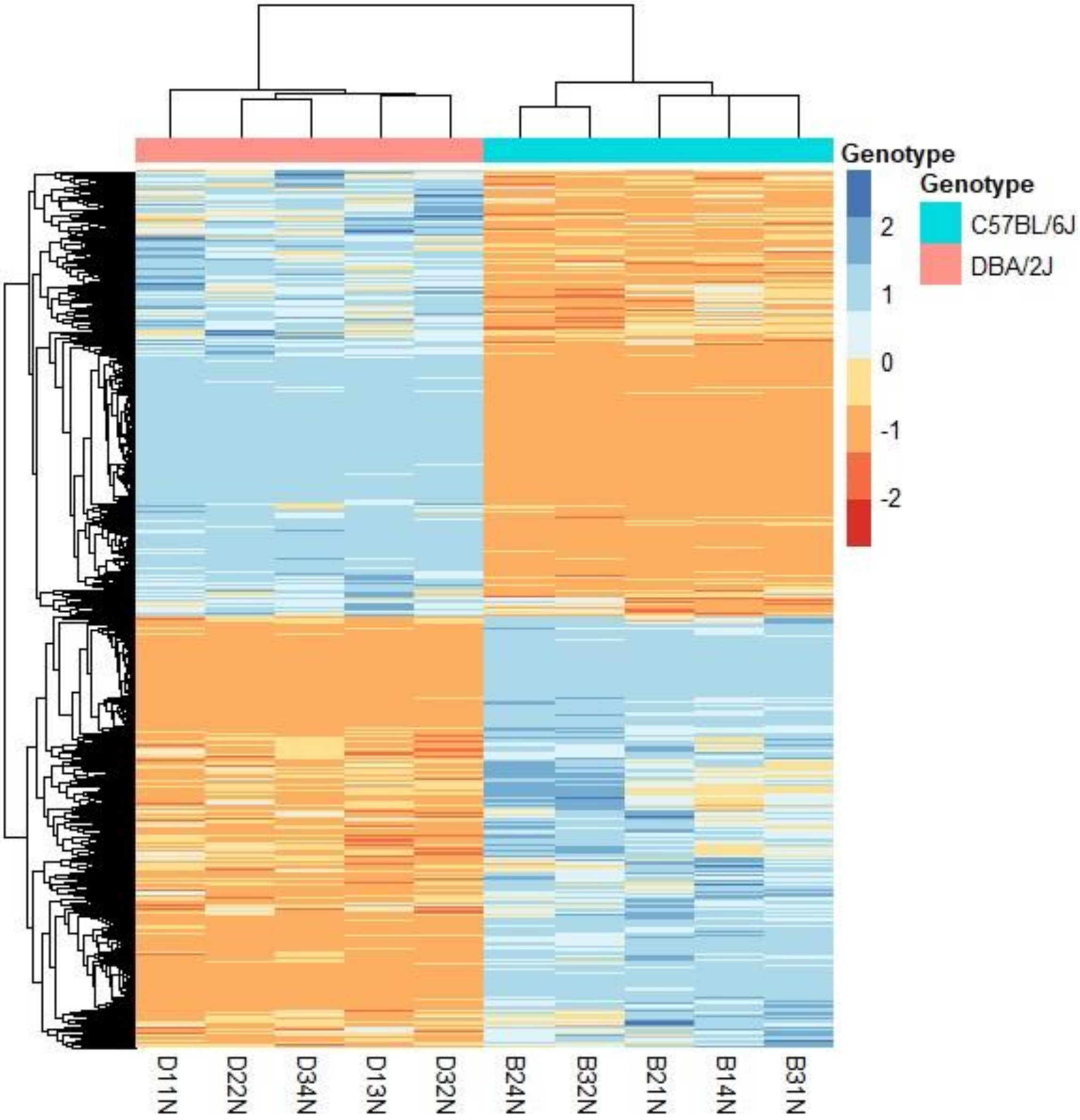

Figure 3.9. Heatmap of hierarchical cluster analysis of differentially expressed genes between B6 aligned B6 and D2 aligned D2 and the log2fold changes (LFCs) of each gene. A positive LFC (Blue) indicates higher expression in D2. The 2-dimensional cluster analysis reveals robust consistency across the samples for differential expression analysis.

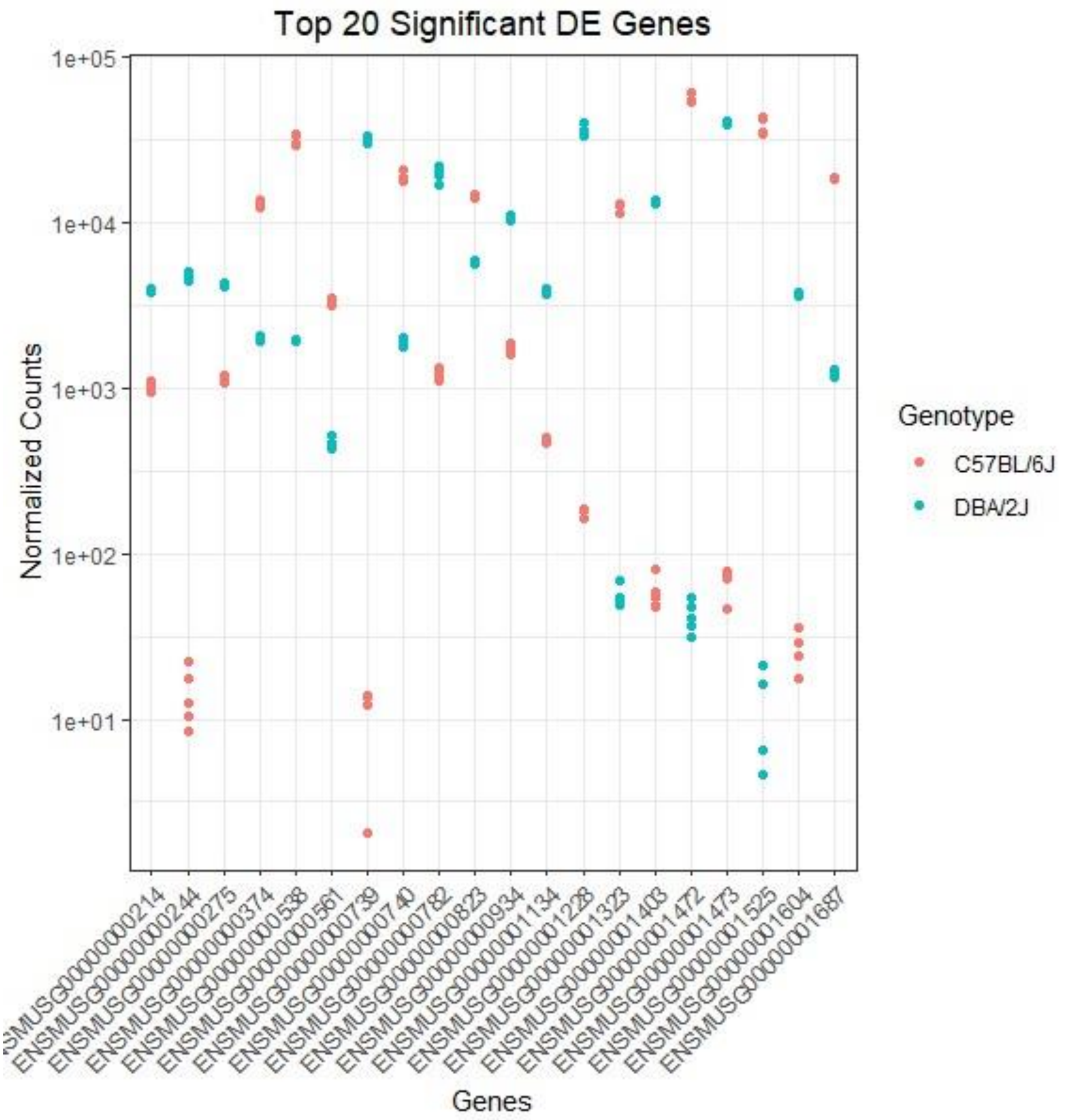

Figure 3.10. Top 20 significantly differentially expressed genes and their normalized counts in each from the D2 aligned analysis.

Comparison of Results

The comparison of results from the B6 aligned D2 analysis and the D2 aligned D2 analysis has been sorted into groups containing only significantly differentially expressed genes. These genes were further sorted by LFC, with positive and negative Log2Fold being compared separately. This is because while the overall LFC of the D2 aligned analysis was significantly higher (B6 aligned D2 = 0.0274, D2 aligned D2 = 0.3270, p < 0.0001), the QC performed showed that the distribution was still even (Figure 3.8). With an even distribution of positive and negative LFCs, the two sets of positive and two sets of negatives were compared to each other to better illustrate the differences between the two comparisons. Positive LFCs indicate increased expression in D2 mice. These analyses have two parts, the comparison of which genes are differentially expressed in each analysis focusing on unique differentially expressed genes, and the comparison of overall LFCs.

The comparison of the negative LFCs showed that 85.05% (2770/3257) of the genes that were significantly differentially expressed in the analysis using B6 aligned D2 samples were also differentially expressed in the analysis using D2 aligned D2 samples (Figure 3.11). The analysis using D2 aligned D2 samples showed significantly more unique genes being differentially expressed than in the B6 aligned analysis (2544/487). The significantly differentially expressed genes with negative LFCs were significantly different in their average LFC (p <0.0001) with the D2 aligned results have a greater magnitude than the B6 aligned results (D2 aligned average negative LFC = -2.5930, B6 aligned average negative LFC = -0.9810).

The comparison of positive LFCs showed that 89.84% (2316/2934) of the genes that were significantly differentially expressed in the analysis using B2 aligned D2 samples were also differentially expressed in the analysis using D2 aligned D2 samples (Figure 3.12). The analysis using D2 aligned D2 samples showed significantly more unique genes being differentially expressed than in the B6 aligned analysis (2544/487). The significantly differentially expressed genes with positive LFCs were significantly different in their average LFC ($p < 0.0001$) with the D2 aligned results have a greater magnitude than the B6 aligned results (D2 aligned average positive LFC = 3.1672, B6 aligned average positive LFC = 1.1465).

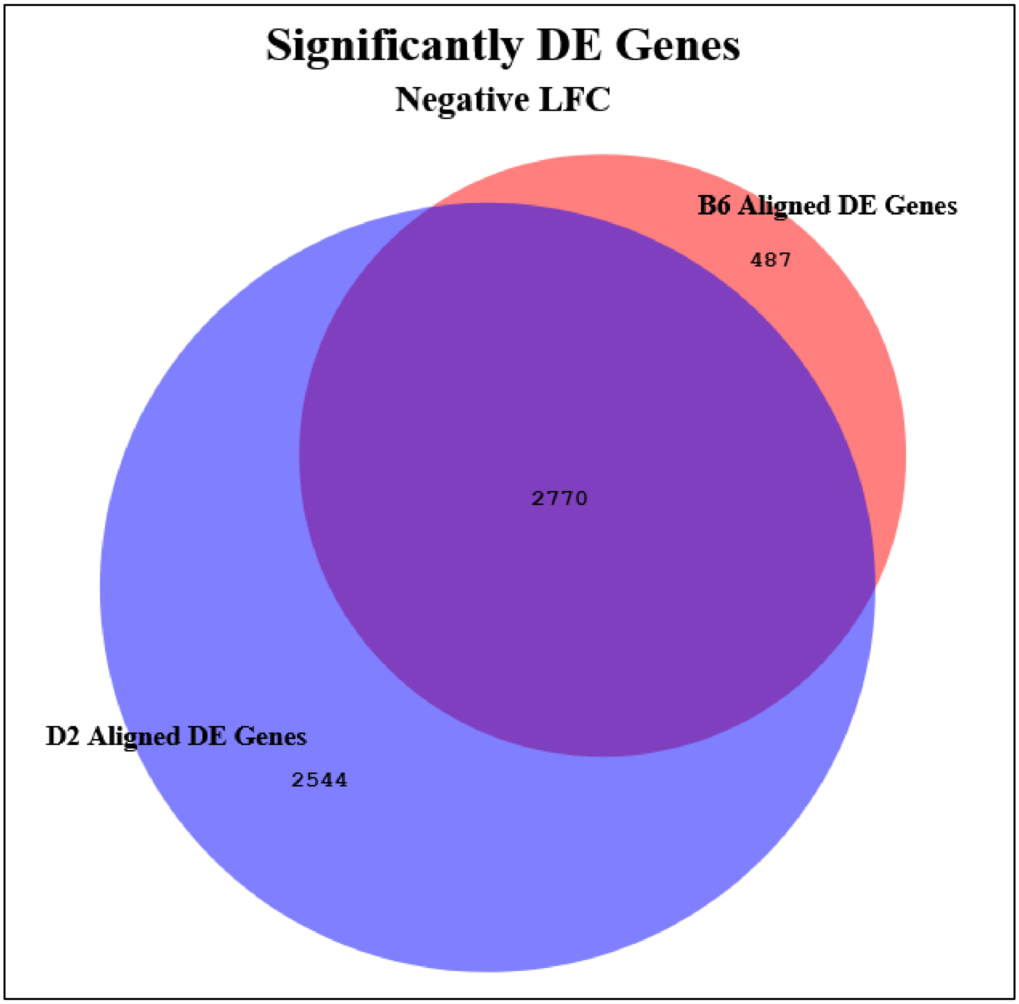

Figure 3.11. Comparison of significantly differentially expressed genes with negative LFCs resulting from differential expression analyses using B6 aligned D2 samples (Red) and D2 aligned D2 samples (Blue). 487 genes were found to be differentially expressed only in the analysis using B6 aligned D2, and 2,544 genes were found to be differentially expressed only in the analysis using D2 aligned D2.

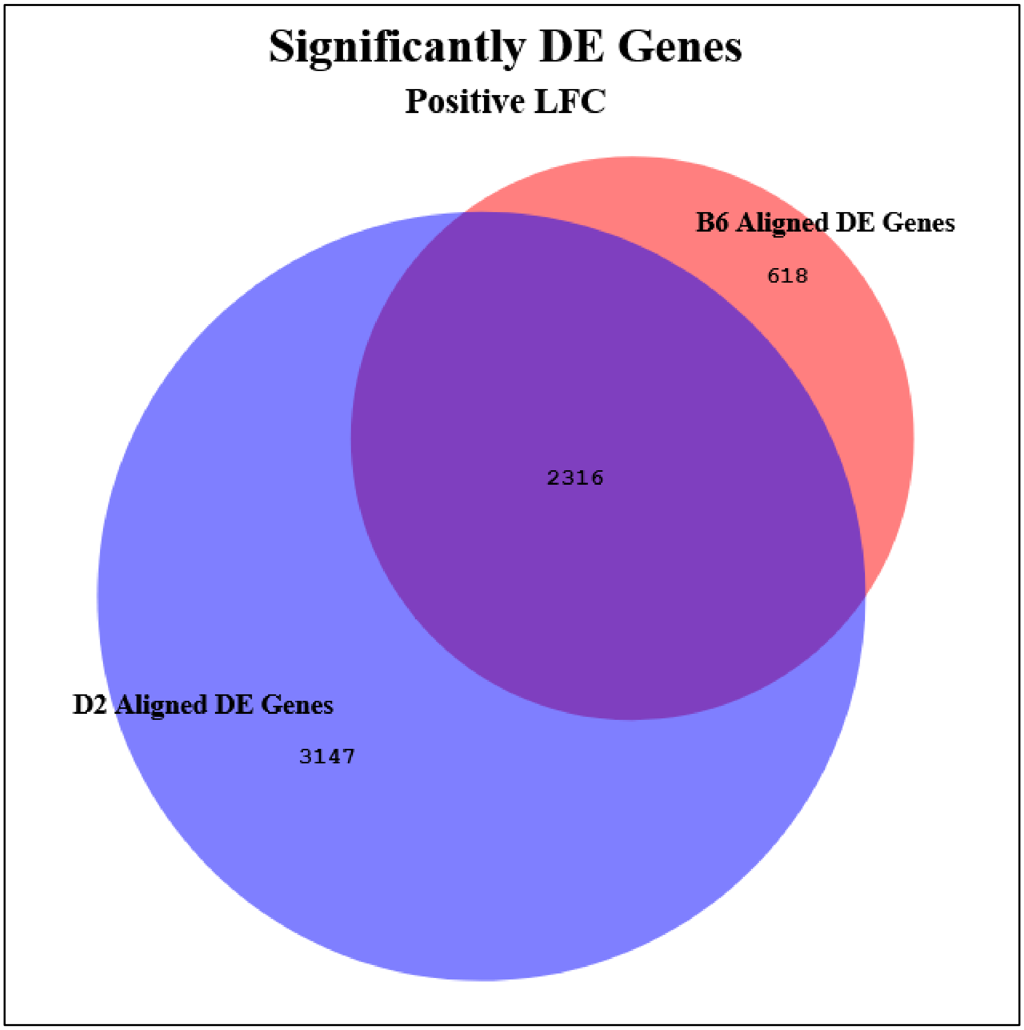

Figure 3.12. Comparison of significantly differentially expressed with positive LFCs genes resulting from differential expression analyses using B6 aligned D2 samples (Red) and D2 aligned D2 samples (Blue). 618 genes were found to be differentially expressed only in the analysis using B6 aligned D2, and 3147 genes were found to be differentially expressed only in the analysis using D2 aligned D2.

Gene Ontology and Semantic Similarity

B6 Aligned D2 Analysis

The gene ontologies of the significantly differentially expressed genes with negative LFCs are shown below, followed by those with positive LFCs. The analysis was run using ToppFun using probability density function to calculate the p value. An FDR correction of 0.05 was used and a gene limit of 3 was set to filter the data. Revigo's semantic similarity analysis was used to cluster and visualize the gene ontology results in the biological processes, molecular function and cellular component categories as scatterplots (Figures 3.13, 3.15, 3.17) and as tree maps (Figures 3.14, 3.16, 3.18). It is important to keep in mind when reading these scatterplots that the axes have no intrinsic meaning. Revigo uses Multidimensional Scaling (MDS) to reduce the dimensionality of a matrix of the GO terms pairwise semantic similarities. This may lead to the result being non-linear, though semantically similar groups will be clustered together. When repeating this analysis, keep in mind that the clusters may appear in different sections of the plot, but the same terms will be clustered together.

B6 Aligned Semantic Similarity of Biological Processes

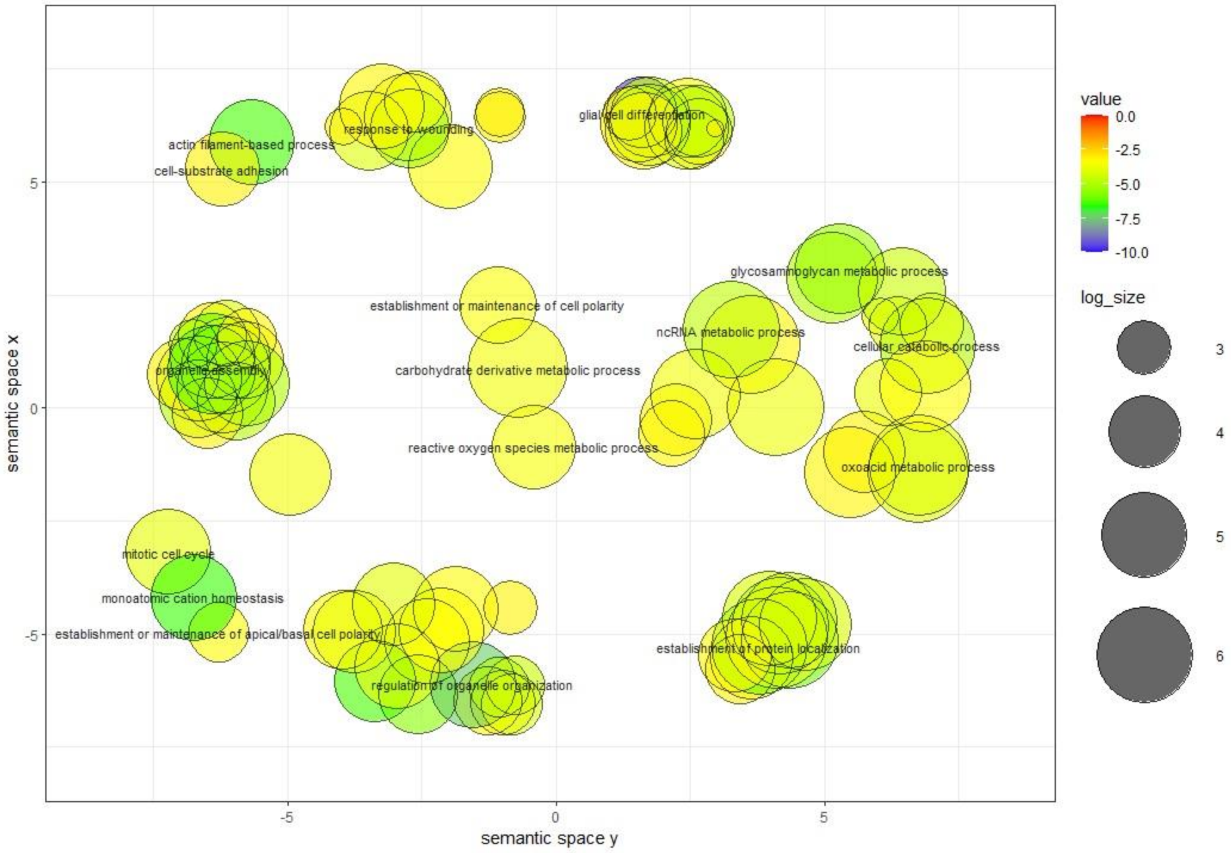

Figure 3.13. Scatterplot of the semantic similarity analysis performed on the biological processes results from the significantly differentially expressed genes from the analysis using B6 aligned D2 samples. Color indicates the log base 10 of the p value output during the ToppFun analysis, with blue indicating the most significantly differentially expressed genes. The size of each point (log_size) indicates the log base 10 of the number of annotations for GO Term ID in selected species in the EBI GOA database.

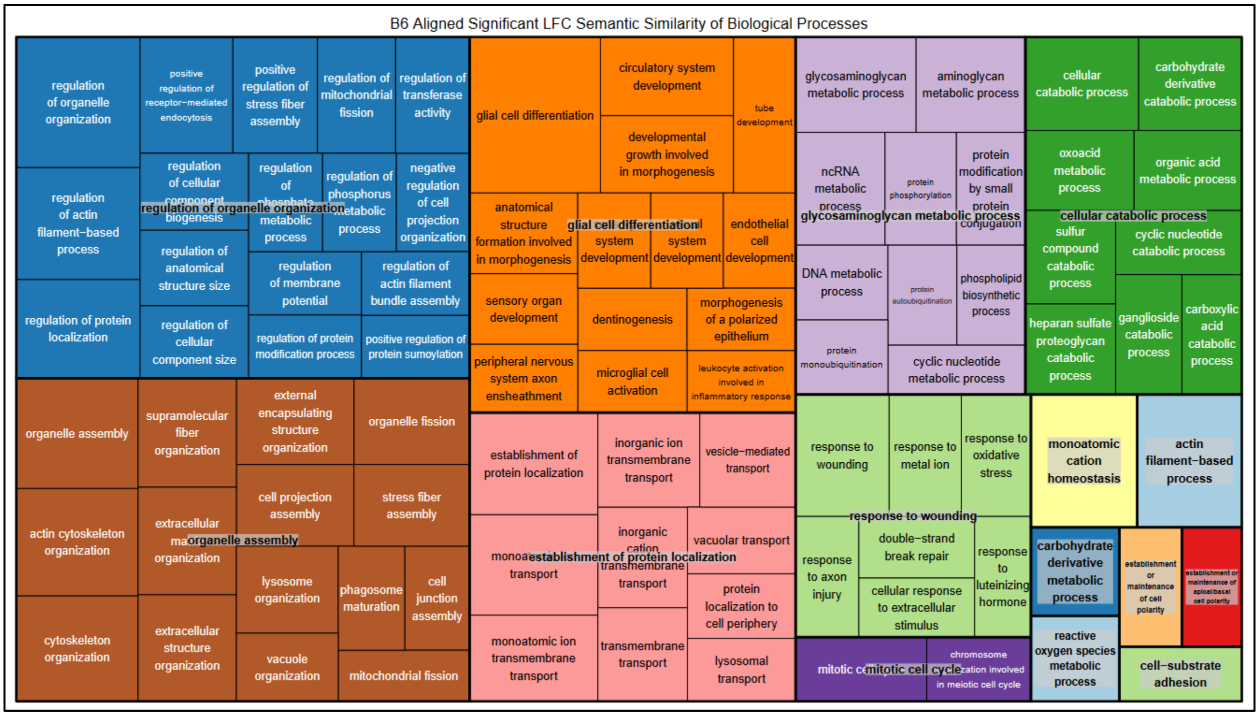

Figure 3.14. Treemap of the semantic similarity analysis performed on the biological processes results from the significantly differentially expressed genes from the analysis using B6 aligned D2 samples. Gene ontology categories are grouped by semantic similarity with closely related categories being clustered together.

B6 Aligned Semantic Similarity of Molecular Functions

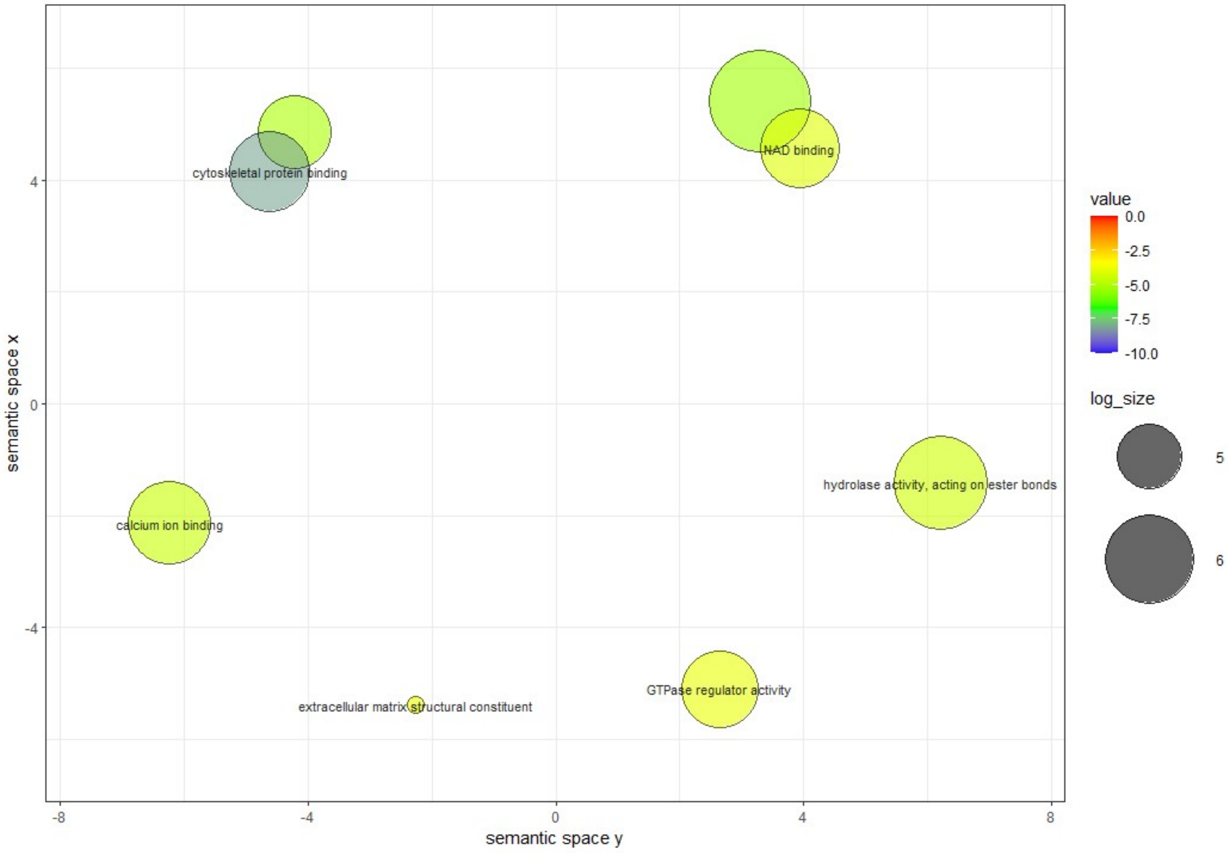

Figure 3.15. Scatterplot of the semantic similarity analysis performed on the molecular function results from the significantly differentially expressed genes from the analysis using B6 aligned D2 samples. Color indicates the log base 10 of the p value output during the ToppFun analysis, with blue indicating the most significantly differentially expressed genes. The size of each point (log_size) indicates the log base 10 of the number of annotations for GO Term ID in selected species in the EBI GOA database.

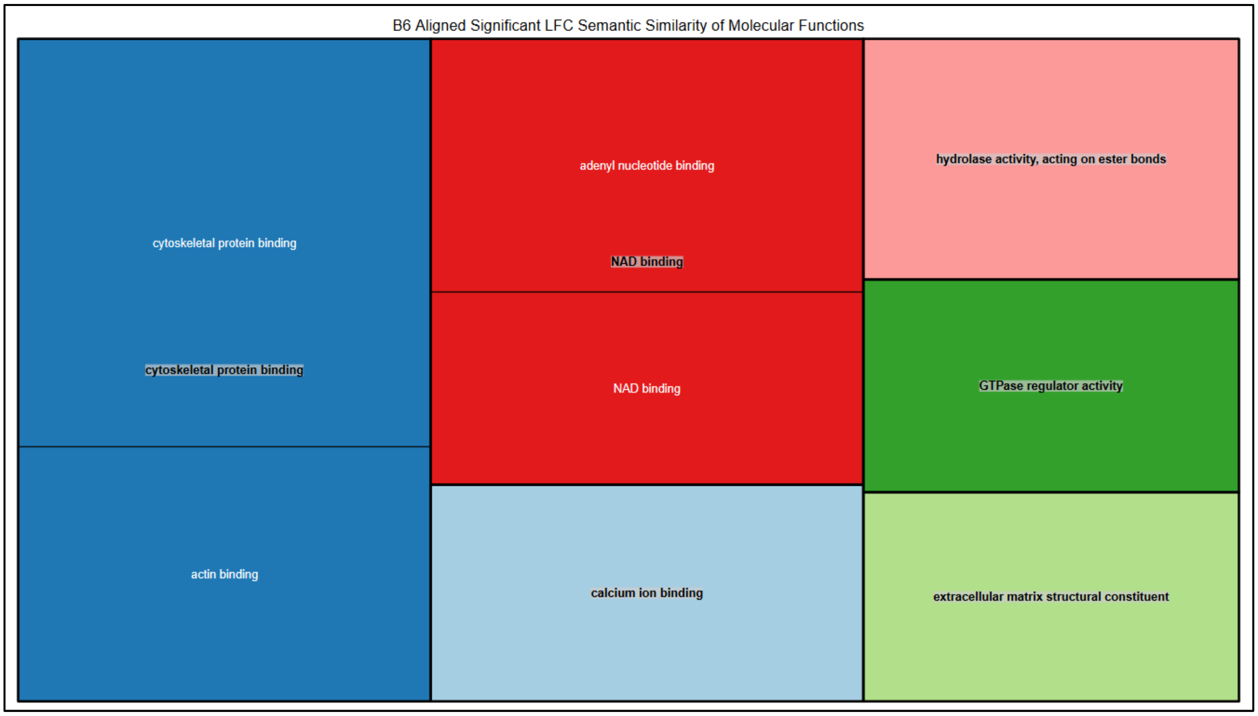

Figure 3.16. Treemap of the semantic similarity analysis performed on the molecular function results from the significantly differentially expressed genes from the analysis using B6 aligned D2 samples. Gene ontology categories are grouped by semantic similarity with closely related categories being clustered together.

B6 Aligned Semantic Similarity of Cellular Components

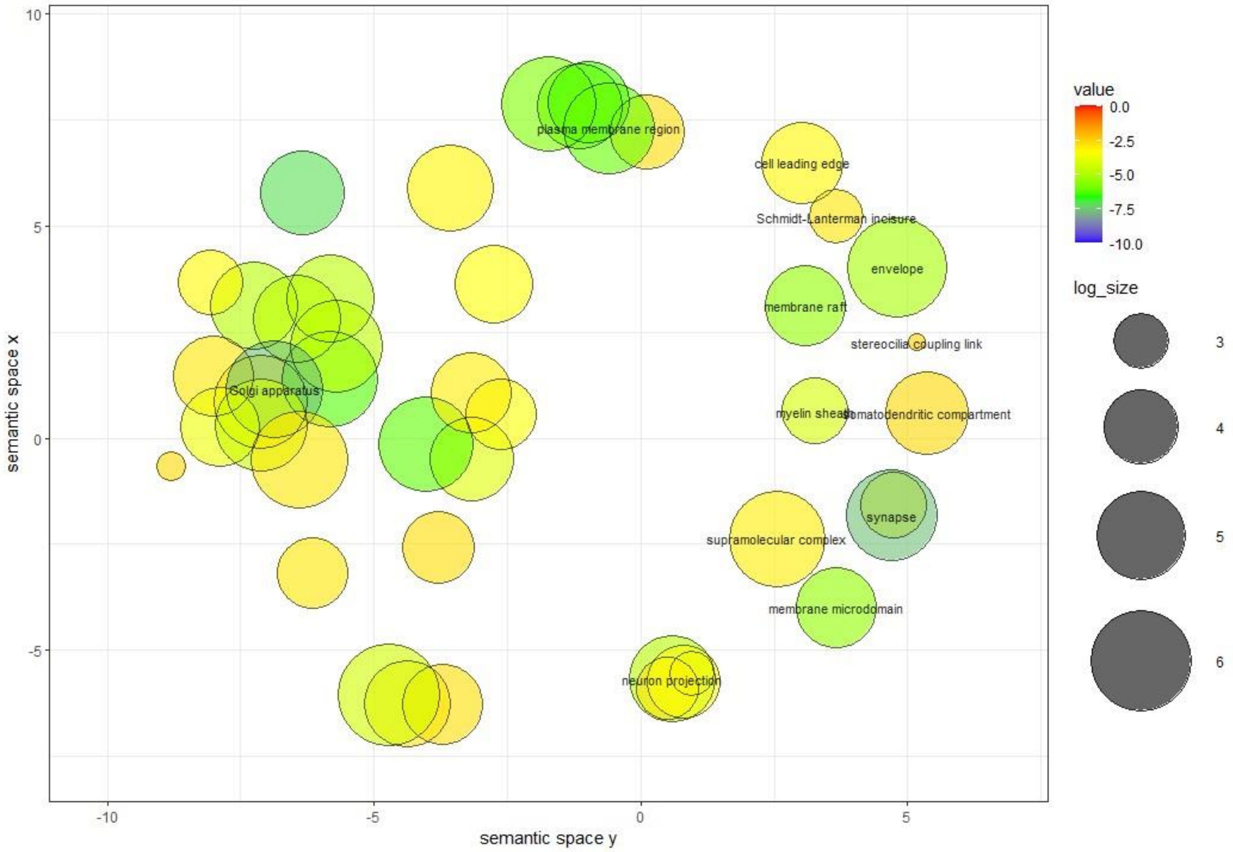

Figure 3.17. Scatterplot of the semantic similarity analysis performed on the cellular component processes results from the significantly differentially expressed genes from the analysis using B6 aligned D2 samples. Color indicates the log base 10 of the p value output during the ToppFun analysis, with blue indicating the most significantly differentially expressed genes. The size of each point (log_size) indicates the log base 10 of the number of annotations for GO Term ID in selected species in the EBI GOA database.

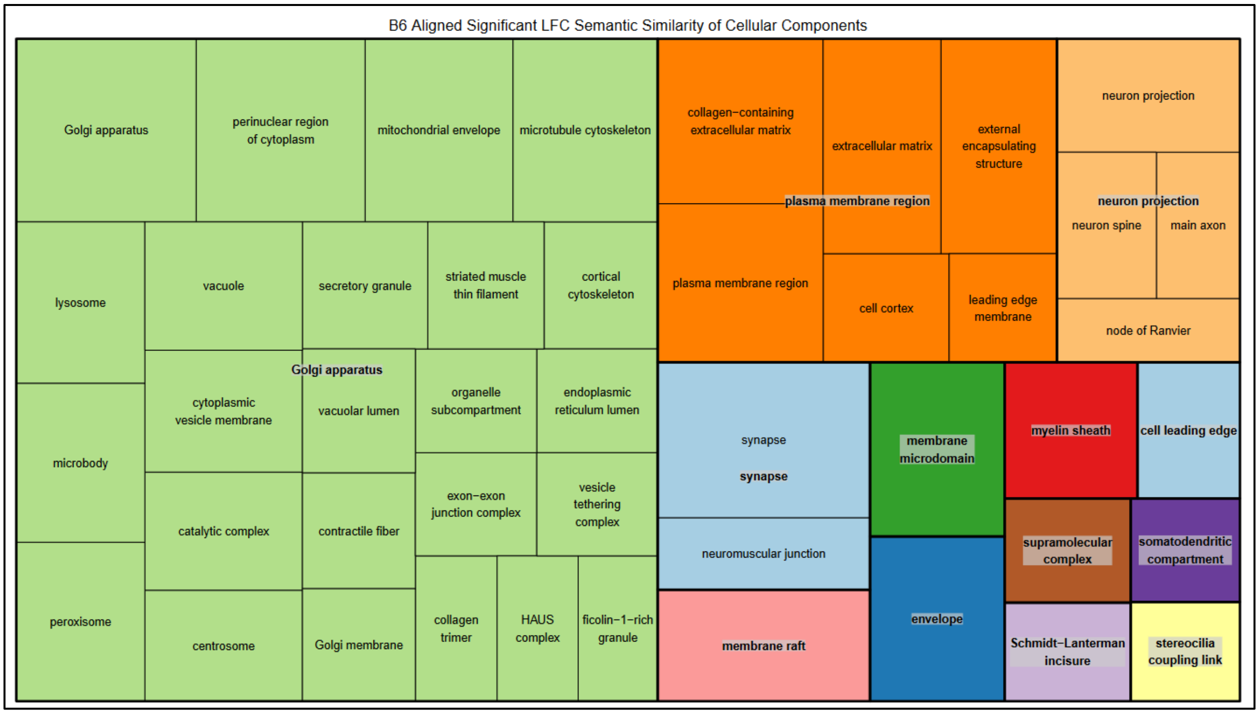

Figure 3.18. Treemap of the semantic similarity analysis performed on the cellular component results from the significantly differentially expressed genes from the analysis using B6 aligned D2 samples. Gene ontology categories are grouped by semantic similarity with closely related categories being clustered together.

D2 Aligned D2 Analysis

The gene ontologies of the significantly differentially expressed genes with negative LFCs are shown below, followed by those with positive LFCs. The analysis was run using ToppFun using probability density function to calculate the p value. An FDR correction of 0.05 was used and a gene limit of 3 was set to filter the data. Revigo's semantic similarity analysis was used to cluster and visualize the gene ontology results in the biological processes, molecular function, and cellular component categories as scatterplots (Figures 3.19, 3.21, 3.23) and as tree maps (Figures 3.20, 3.22. 3.24).

D2 Aligned LFC Semantic Similarity of Biological Processes

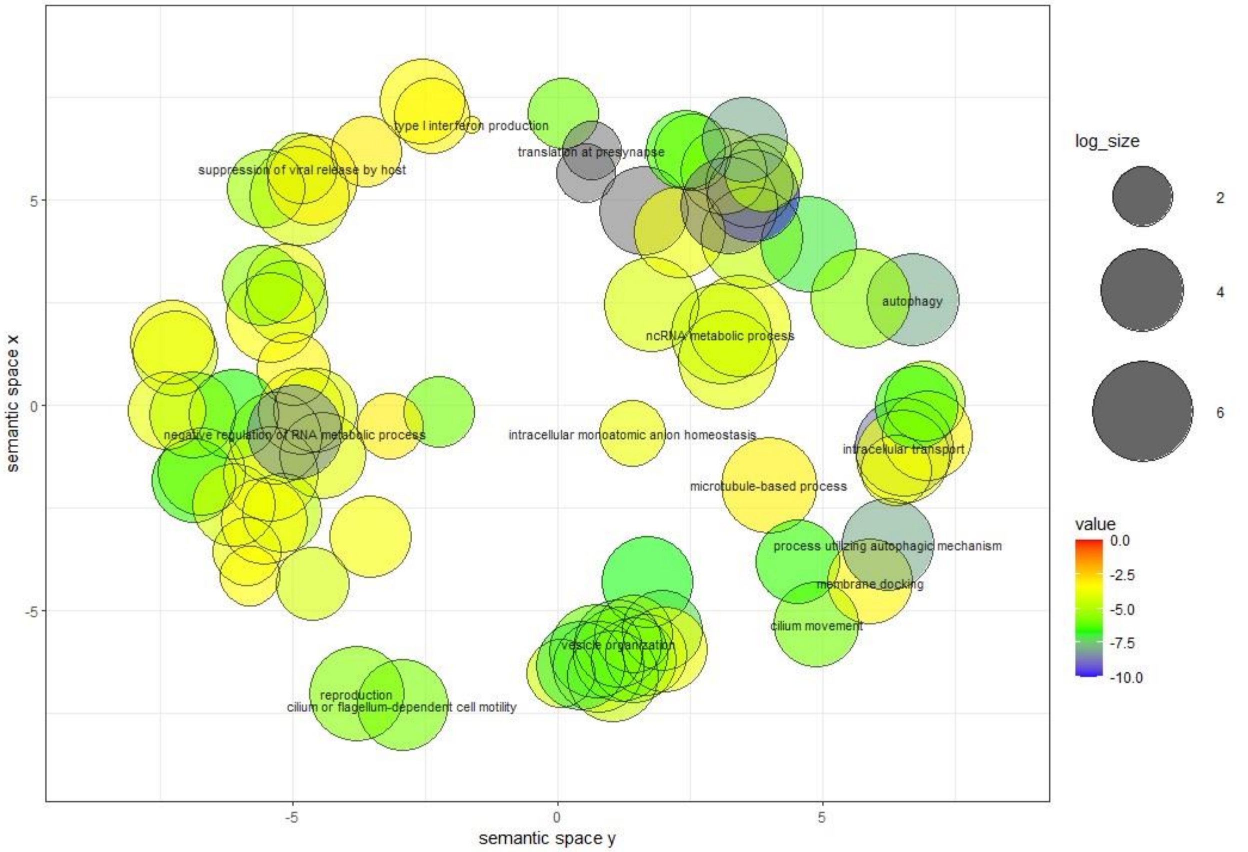

Figure 3.19. Scatterplot of the semantic similarity analysis performed on the biological processes results from the significantly differentially expressed genes from the analysis using D2 aligned D2 samples. Color indicates the log base 10 of the p value output during the ToppFun analysis, with blue indicating the most significantly differentially expressed genes. The size of each point (log_size) indicates the log base 10 of the number of annotations for GO Term ID in selected species in the EBI GOA database.

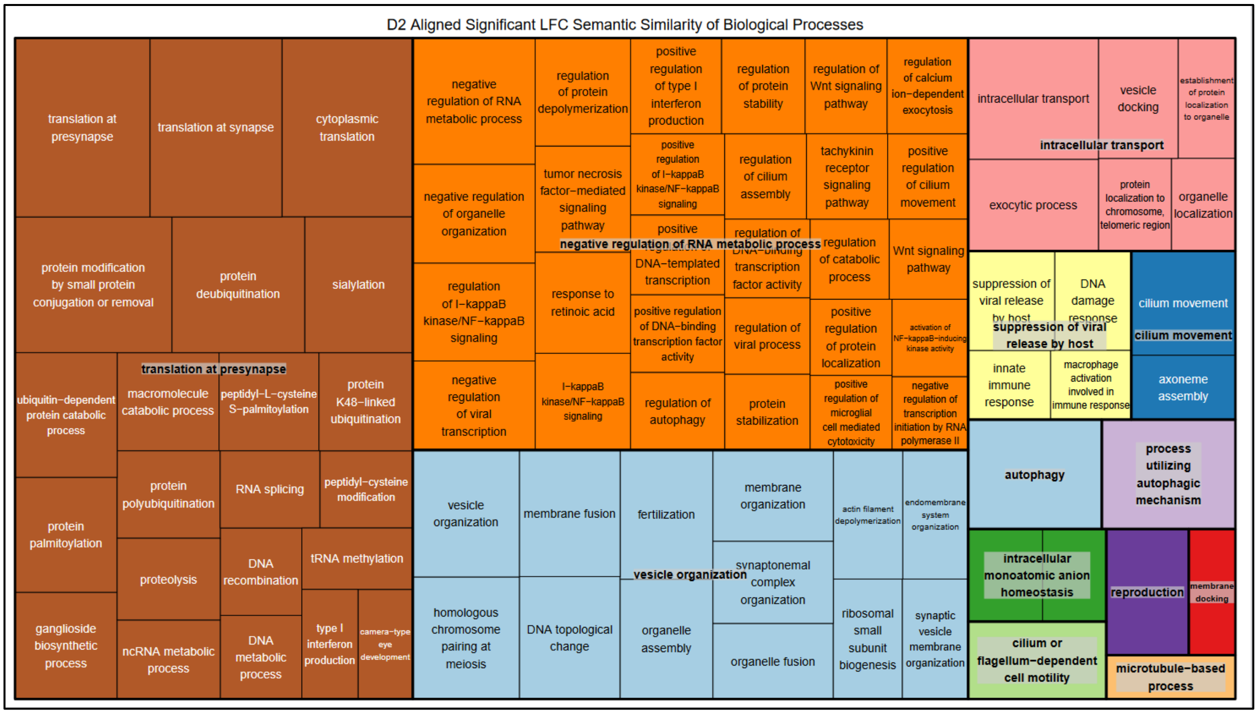

Figure 3.20. Treemap of the semantic similarity analysis performed on the biological processes results from the significantly differentially expressed genes from the analysis using D2 aligned D2 samples. Gene ontology categories are grouped by semantic similarity with closely related categories being clustered together.

D2 Aligned Semantic Similarity of Molecular Functions

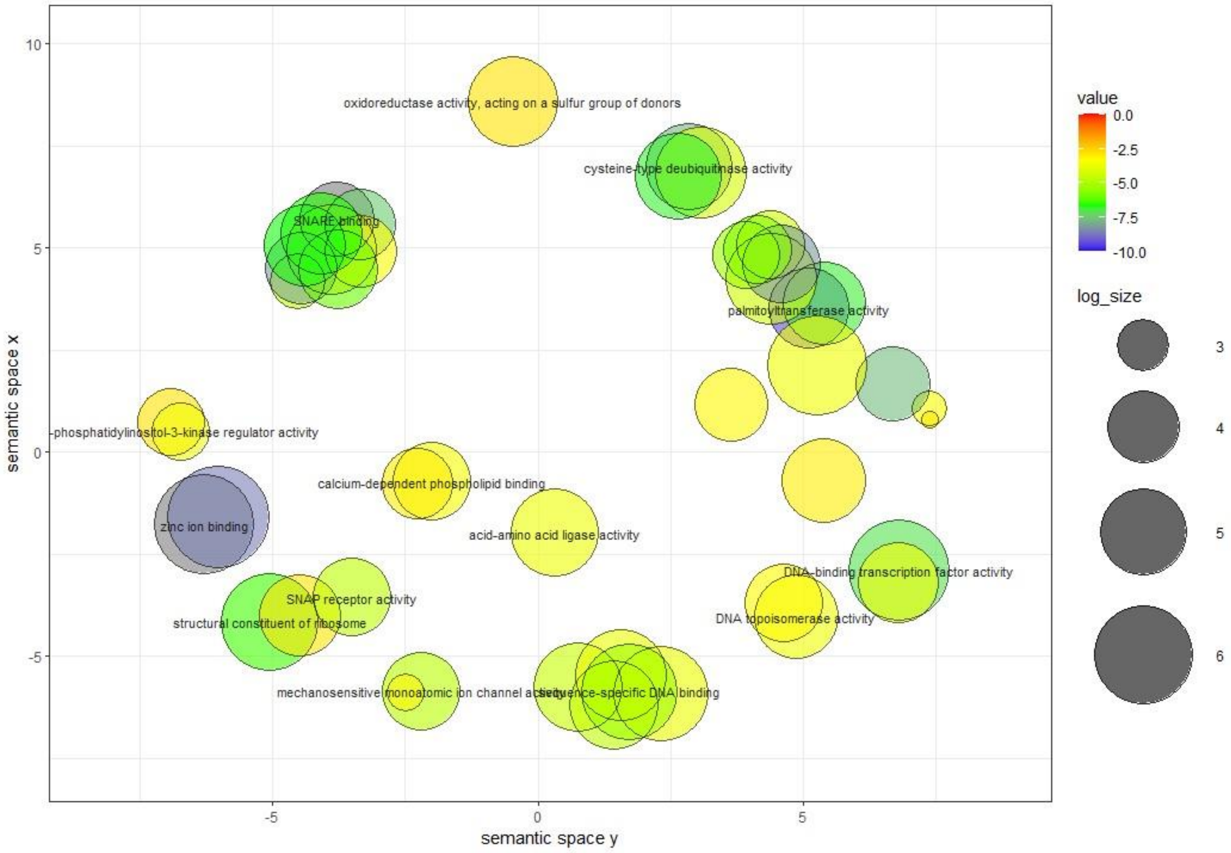

Figure 3.21. Scatterplot of the semantic similarity analysis performed on the molecular function results from the significantly differentially expressed genes from the analysis using D2 aligned D2 samples. Color indicates the log base 10 of the p value output during the ToppFun analysis, with blue indicating the most significantly differentially expressed genes. The size of each point (log_size) indicates the log base 10 of the number of annotations for GO Term ID in selected species in the EBI GOA database.

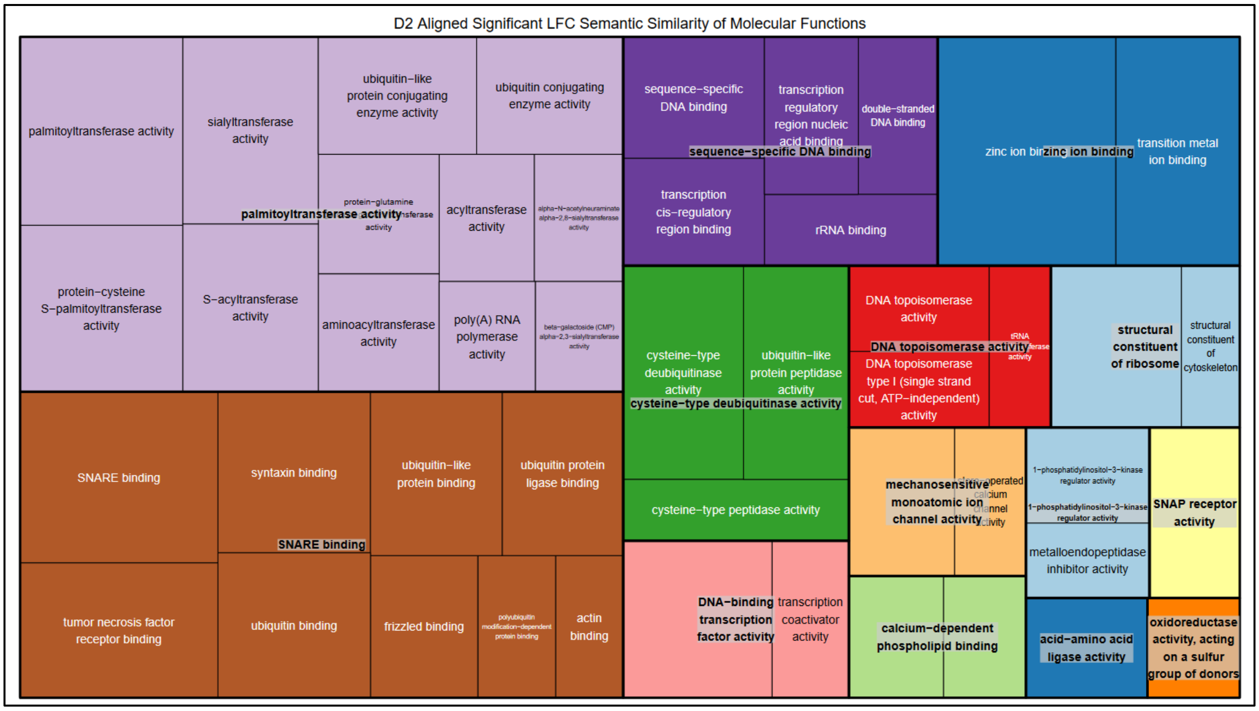

Figure 3.22. Treemap of the semantic similarity analysis performed on the molecular function results from the significantly differentially expressed genes from the analysis using D2 aligned D2 samples. Gene ontology categories are grouped by semantic similarity with closely related categories being clustered together.

D2 Aligned Semantic Similarity of Cellular Components

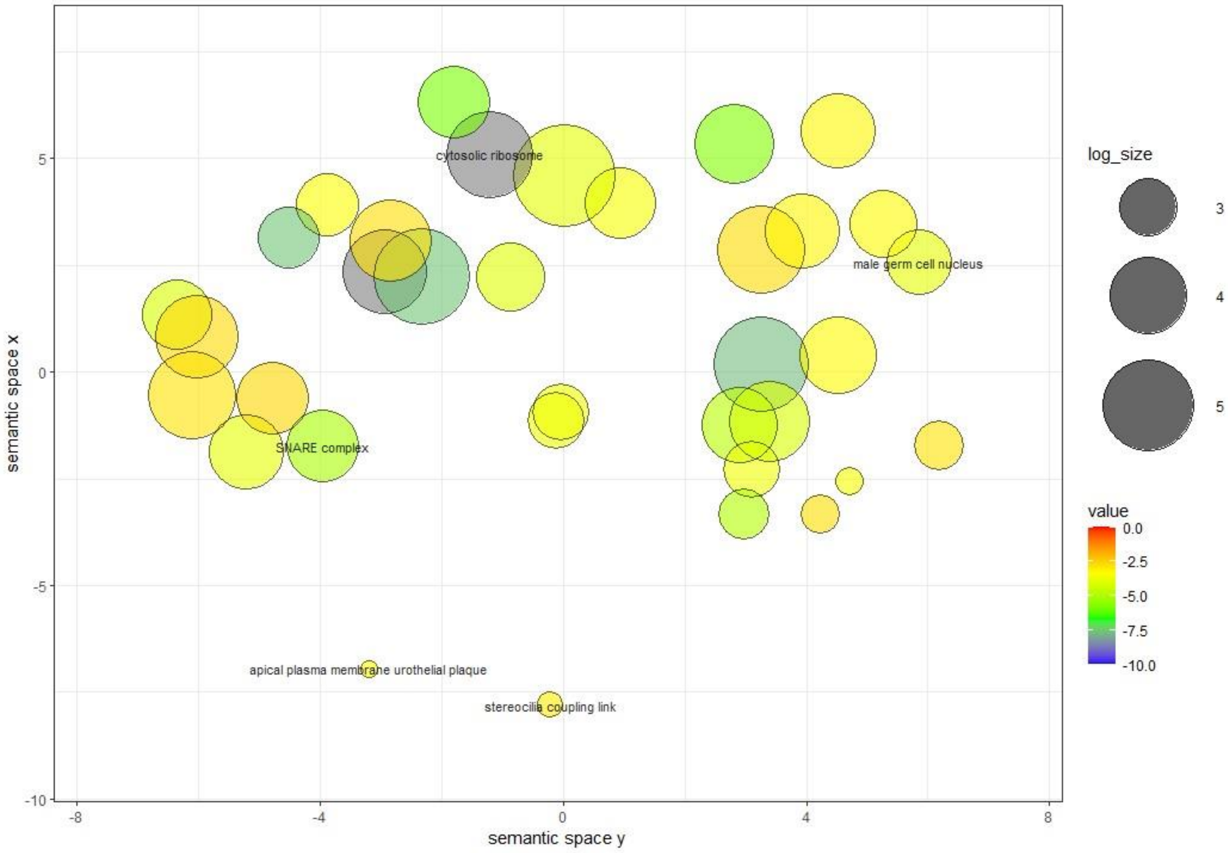

Figure 3.23. Scatterplot of the semantic similarity analysis performed on the cellular component results from the significantly differentially expressed genes from the analysis using D2 aligned D2 samples. Color indicates the log base 10 of the p value output during the ToppFun analysis, with blue indicating the most significantly differentially expressed genes. The size of each point (log_size) indicates the log base 10 of the number of annotations for GO Term ID in selected species in the EBI GOA database.

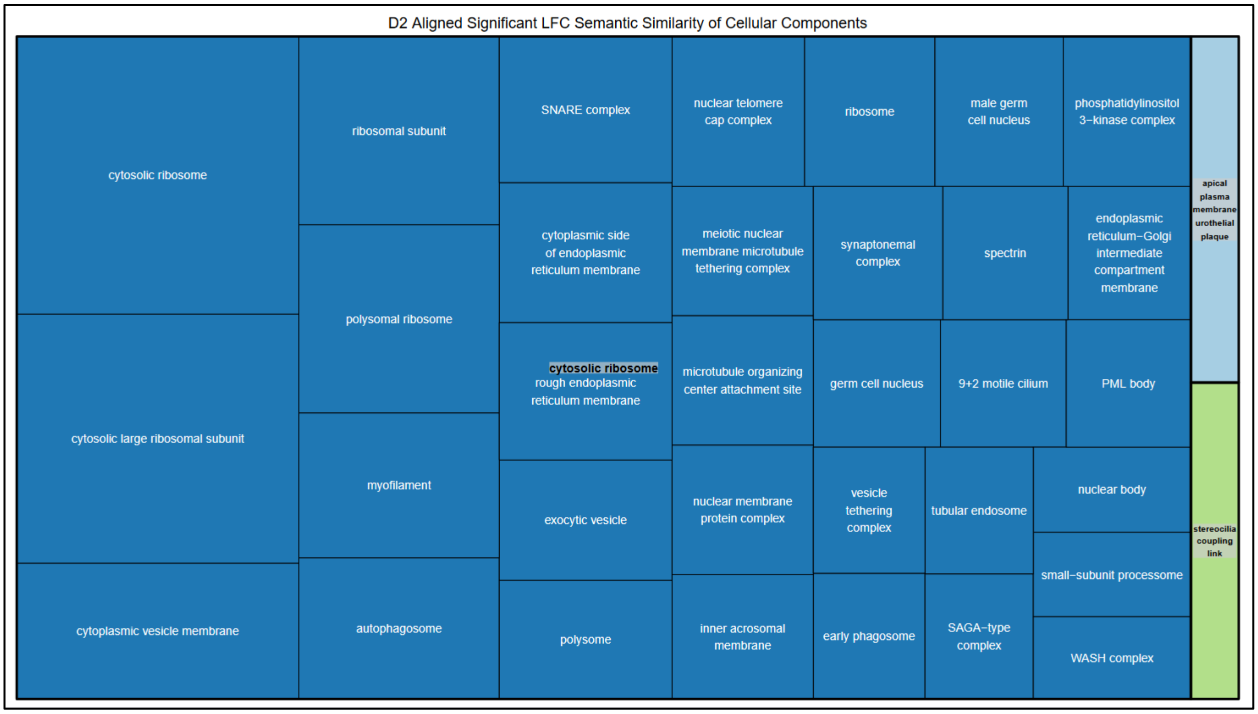

Figure 3.24. Treemap of the semantic similarity analysis performed on the cellular component results from the significantly differentially expressed genes from the analysis using D2 aligned D2 samples. Gene ontology categories are grouped by semantic similarity with closely related categories being clustered together.

Discussion

Alignment

Aligning D2 reads to the D2 genome produced a higher alignment percentage than aligning them to the B6 genome (Table 2). This will lead to more accurate gene quantification, increased sensitivity of the analysis with respect to genes and transcripts with low expression, and a reduction in background noise from unaligned reads (Oshlack et al., 2010)**.** This again is promising, as it serves to show that there is a benefit to aligning to the D2 reference genome, though determining the scope and scale of that benefit is still in progress.

Differential expression and comparison of results

Results of the differential expression analyses have shown that the differential expression between B6 aligned B6 and B6 aligned D2 is significant (Figure 3.5.) This is in keeping with the results found by Putman et al. (2016) and Kearns et al. (2005), and show that the basis for this study is well founded. By analyzing the resulting LFCs, there was a lack of bias in expression direction, with similar numbers of genes having positive and negative LFCs (Figures 3.11 and 3.12). The top 20 most significantly differentially expressed genes showed this same trend, indicating that the analysis was run correctly. If there was a significant bias towards positive or negative LFC values, then that would indicate a problem either in sample preparation, leading to one set of mice to have consistently higher or lower gene expression, or a problem in the analysis itself such as during normalization.

The analysis run using D2 aligned D2 samples showed similar quality control metrics (Figures 3.7-3.8) as the B6 aligned (Figures 3.2-3.4). There was a lack of bias in expression direction observed in the resulting LFCs, and the top 20 most significantly differentially expressed genes again showing no bias towards positive or negative LFC (Figure 3.10). This indicates that the D2 aligned D2 samples are meeting the same quality control metrics as the B6 aligned D2 samples, and the comparison between the two sets of results can be done with confidence in the preparation and setup of the analyses.

The analysis run using D2 aligned D2 samples found significantly more significantly differentially expressed genes (p < 0.05) than the analysis run using the B6 aligned D2 samples, with 10,778 differentially expressed genes found in the D2 aligned D2 sample compared to 6,191 in the B6 aligned D2 analysis. In both the positive and negative LFCs a large amount of overlap was seen between the results of the two analyses (Figures 3.11 and 3.12). However, the D2 aligned results showed significantly more uniquely differentially expressed genes in both positive and negative directions (3,147 and 2,544 respectively). The D2 aligned average negative LFC was -2.5930 and the B6 aligned average negative LFC was -0.9810. The D2 aligned average positive LFC was 3.1672, and B6 aligned average positive LFC was 1.1465 This, along with the very high overlap of the B6 aligned results, where 85.05% (Negative LFC) and 89.84% (Positive LFC) were seen to be differentially expressed in the same direction as the D2 aligned results, shows that when aligning to the D2 genome there is a significant increase in the total number of differentially expressed genes while retaining a majority of the differentially expressed genes seen in the B6 aligned analysis.

This indicates that aligning to the D2 genome provides a 74.26% increase in significantly differentially expressed genes with negative LFCs and a 109.20% increase in significantly differentially expressed genes with positive LFCs. The total increase in significantly differentially expressed genes identified is 90.17%.

Gene ontology and semantic similarity

The gene ontology categories across both analyses showed similar levels of significance, with most sitting between -2.5 and -5 on a log10 scale of the p values. There were a few outliers, but the outliers were more significant than the average, with no low significance outliers. This indicates that the gene ontology categories can be considered reliable indications of the functions of the genes involved.

When comparing the categories between the B6 aligned analysis and the D2 aligned analysis, there is a low amount of overlap in the semantic clusters. There were overlaps, for example in the biological processes, gliogenesis showed the same amount of differentially expressed genes in both analyses. However, none of the other categories overlapped, which is likely at least partially caused by the large increase in differentially expressed genes in the D2 aligned analysis. This theme continues throughout the analyses, with there being some overlap but not overlap completely. One important category seen in the D2 aligned results that was not seen in the B6 aligned data was ribosomal protein gene expression. Categories such as "translation at synapse", "translation at postynapse", etc. This suggests that the D2 RNA-seq data shows a decrease in expression of the expression of ribosomal mRNAs when analyzed with the D2 alignment. Decreased ribosomal protein expression reflects a decreased capacity for protein translation, so this could be a biologically relevant difference discovered by aligning to the D2 genome that would not have been identified while aligning to the B6 genome.

Chapter 4: Differential Exon Utilization and Alternative Splicing

Introduction

Differential exon utilization occurs when exons within a gene are included in or excluded from the final RNA transcripts produced by a cell when compared between two experimental groups. DEXSeq (Anders et al., 2012) is an R/Bioconductor package that uses a statistical method to test for differential exon usage in RNA-seq data. It uses generalized linear models to do this, and takes biological variation into account to control false discoveries. It also identifies differences in splicing and translation.

DEXSeq requires exon count data which is prepared using scripts provided in the package. To generate this count data, RNA-Seq data, a genome fasta file, and an annotation GTF file are required. Indexed BAM files are generated first, using the fasta file and the annotation, then the annotation needs to be flattened for use with DEXSeq. These exon counts are then used in DEXSeq's analysis of differential exon utilization.

Gene ontology can be run on the genes with differentially utilized exons, providing an understanding of their functions. These GO categories reflect the functions of the genes with differentially utilized exons, and when comparing results of multiple analyses Revigo can be used to reduce the data and make visualization easier.

This study is running two different DEXSeq analyses, one using B6 aligned D2 samples and one using D2 aligned D2 samples. This will allow for a comparison using Welch's t-tests and direct list comparison of the genes with differentially utilized exons, as well as a comparison of the exons. The LFCs of both analyses will be used to compare the magnitude of the differentially expressed exons, broken in to positive and negative LFC groups.

Methods

Count Data Preparation

The RNA-Seq data was prepared for DexSeq analysis using the VCU Group Server on the VIPBG Cluster System First the GTF file converted in the previous chapter (See chapter 2) needed to be flattened for use with DexSeq (Bioconductor) (Anders et al., 2012). The script dexseq_prepare_annotation.py (Appendix 2) requires a GTF file as input. However, the process of converting the GFF3 file to a GTF file left certain attributes that were not compatible with the conversion script. With the aid of Dr. Mikhail Dozmorov, the specific attribute (parent) causing errors was identified and removed. The parent attribute is part of GFF3 files, indicating the parent transcript for each entry. It is not present in GTF files, and was not removed by AGAT in the conversion to GTF. Removing it does not change the function of the annotation file, but allows it to function in the dexseq_prepare_annotation.py. The resulting GTF file (DBA_2J_v3.2_3_14_23_filtered.gtf) was then run through the aforementioned script, resulting in a flattened GFF file (DBA_2K_v3.2_flattened.gff). This flattened file was then used with the script dexseq_count.py, to generate the dexseq counts. The BASH script used to run this on the VCU group server was DexSeqCounts.sh. The B6 aligned counts (both B6 and D2) were previously prepared using the same sample data by Emma Gnatowski.

DexSeq Analysis

DexSeq was run using R version 4.3.1 (DEXSeq_Analysis_script.R) on the B6 aligned B6 and the B6 aligned D2 counts, and the significant results (FDR of 0.1) analyzed for expression patterns and gene ontology. This was repeated using B6 aligned B6 counts and D2 aligned D2 counts.

Gene Ontology Analysis

The gene ontology of the genes showing differential exon utilization was run using ToppFun using a probability density function to calculate p values, then filtering for a false discovery rate of 0.05 and gene limits of 3 or more.

Comparison of Results

The results of both DexSeq analyses were then compared to each other using direct comparisons to determine the total number of differentially utilized exons and differentially expressed genes, and t-tests to determine differences in the magnitude of those changes, measured by LFC. The LFC comparison was broken down into positive and negative LFC groups, as the overall average for both analyses was nearly zero. The gene ontology categories were compared using Revigo to reduce the number of categories and cluster by semantic terms. Lastly, specific genes of import were taken to use as examples of how aligning to the D2 genome can improve differential exon utilization results. The genes chosen for this were Ninein, Gabra2, and Gsk3b. These genes were chosen both for their import in ongoing AUD research and for how the D2 alignment affected them.

Results

Count Data Preparation

The GTF and count data files were successfully prepared, leading to the generation of count data (Appendix 1, B6 Aligned DexSeq Counts, D2 Aligned DexSeq Counts).

B6 Aligned DexSeq Analysis

21,223 significantly ($p < 0.05$, FDR 0.1) differentially utilized exons were identified in the B6 aligned DexSeq analysis. There were 6,650 genes with differentially utilized exons. There were 10,566 exons with positive LFC indicating higher expression in D2, and 10,652 with negative LFC. These exons had an average positive LFC of 0.8323 and an average negative LFC of -1.2138.

B6 Aligned Analysis Hierarchical Heatmap of Correlation Data

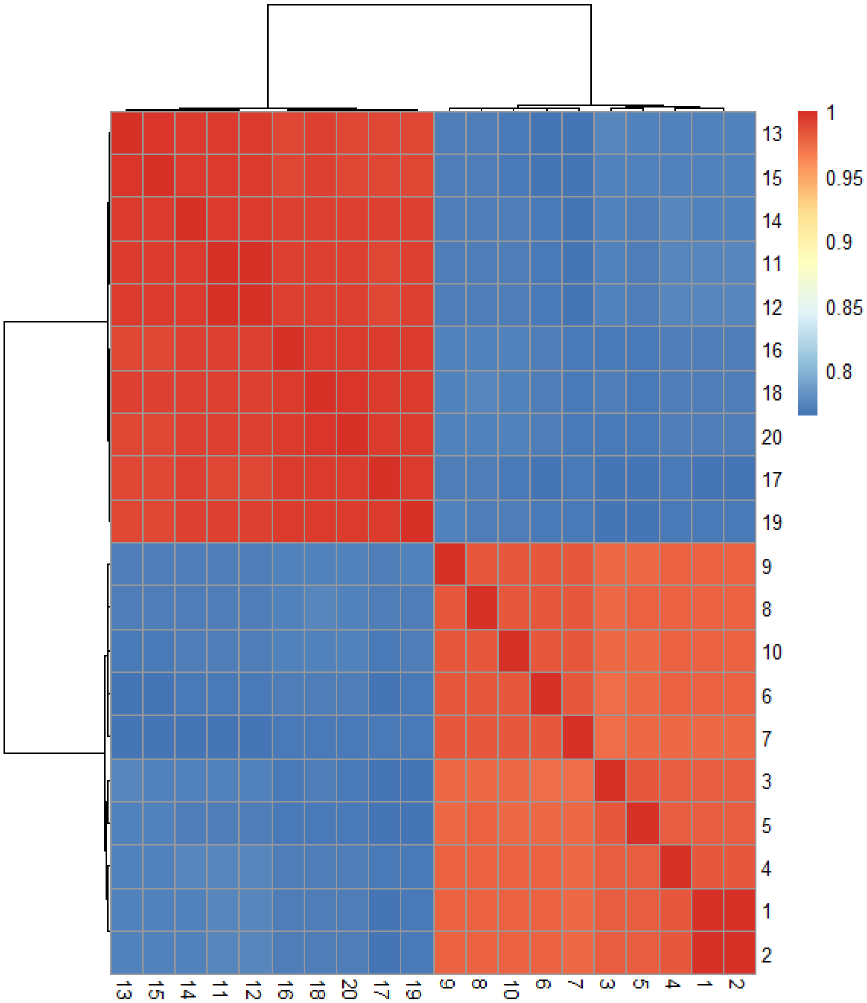

Figure 4.1. Heatmap of B6 aligned D2 correlation data. There are 10 samples used as input, but 20 columns used in this analysis. The first 10 correspond to the number of reads mapping to out exonic regions, and the last 10 correspond to the sum of the counts mapping the rest of the exons from the same gene on each sample. Samples 1-5 are B6 counts, specifically B14, B21, B24, B31, and B32, and samples 6-10 are D2 counts, specifically D11, D13, D22, D32, and D34. 11-15 are B6, the same samples, and 16-20 are D2 aligned, the same samples.

B6 Aligned Analysis LFC vs Mean Expression

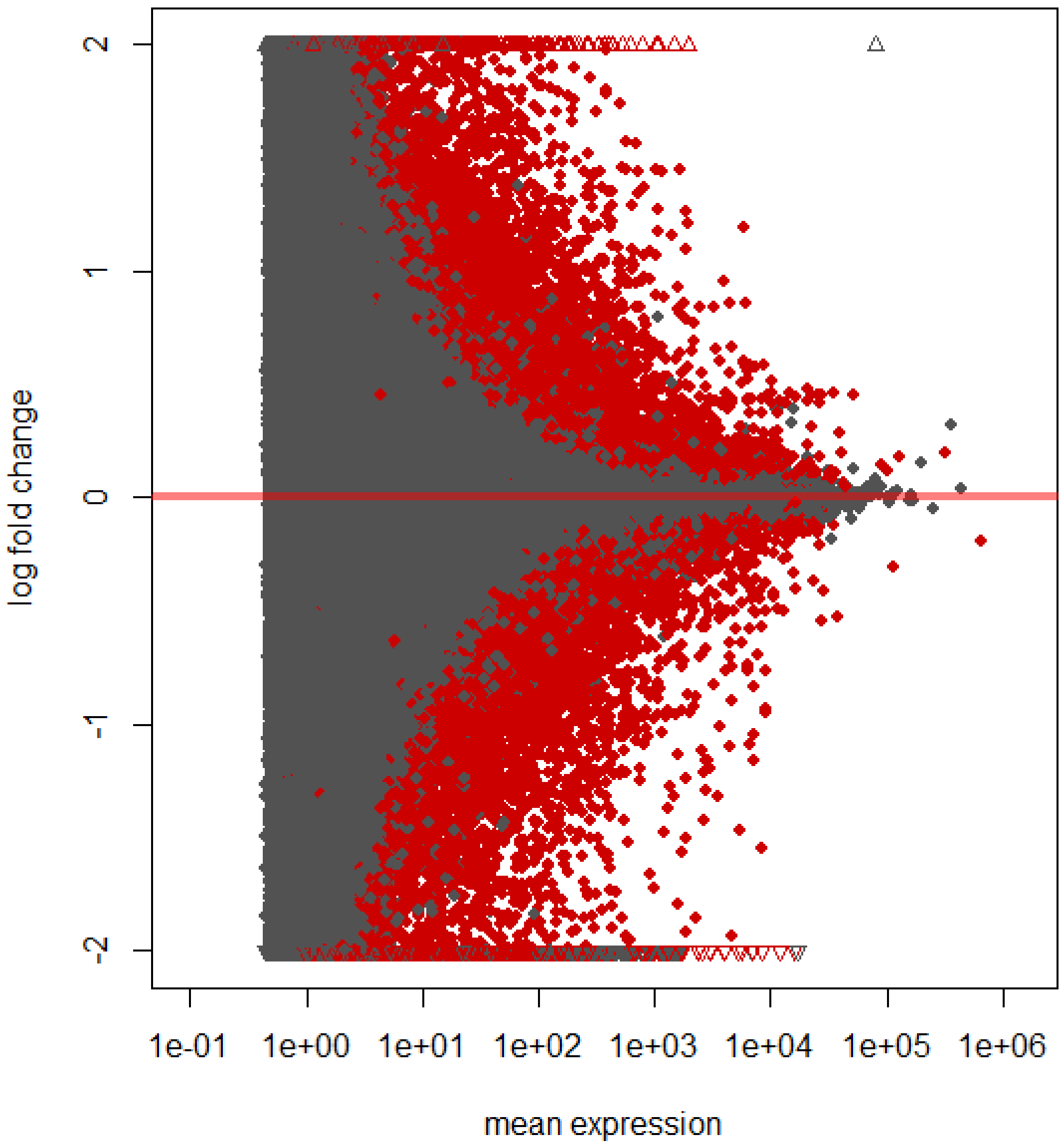

Figure 4.2. LFC of differentially expressed exons compared to mean expression for the analysis using B6 aligned D2 samples. Red indicates significantly differentially expressed genes.

D2 Aligned DexSeq Analysis

81,206 significantly ($p < 0.05$, FDR 0.1) differentially utilized exons were identified in the D2 aligned DexSeq analysis. There were 13,521 genes with differentially utilized exons. 43,775 of the exons had a positive LFC and 37,089 had a negative LFC. The average positive LFC was 1.5800 and the average negative LFC was -1.3011.

D2 Aligned Analysis Hierarchical Heatmap of Correlation Data

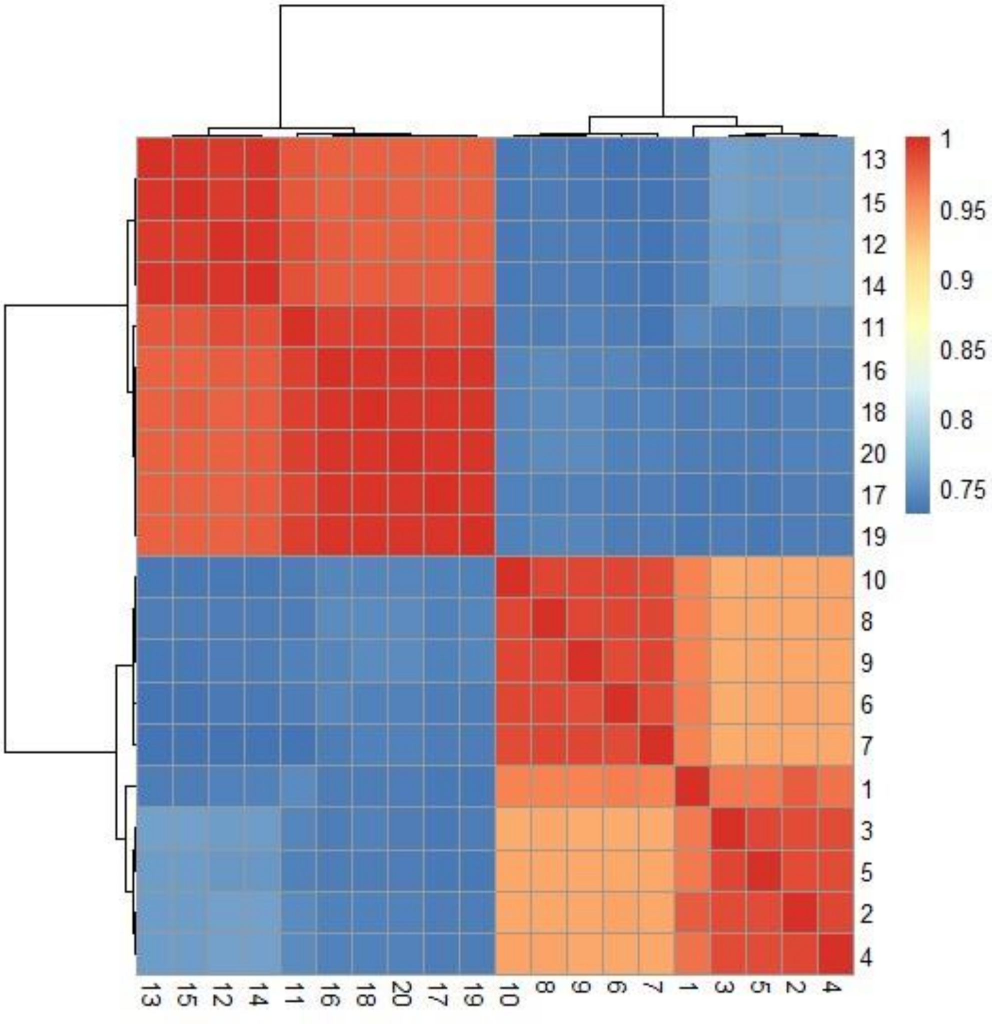

Figure 4.3. Heatmap of D2 aligned D2 correlation data. There are 10 samples used as input, but 20 columns used in this analysis. The first 10 correspond to the number of reads mapping to out exonic regions, and the last 10 correspond to the sum of the counts mapping the rest of the exons from the same gene on each sample. Samples 1-5 are B6 counts, specifically B14, B21, B24, B31, and B32, and samples 6-10 are D2 counts, specifically D11, D13, D22, D32, and D34. 11-15 are B6, the same samples, and 16-20 are D2 aligned, the same samples.

D2 Aligned Analysis LFC vs Mean Expression

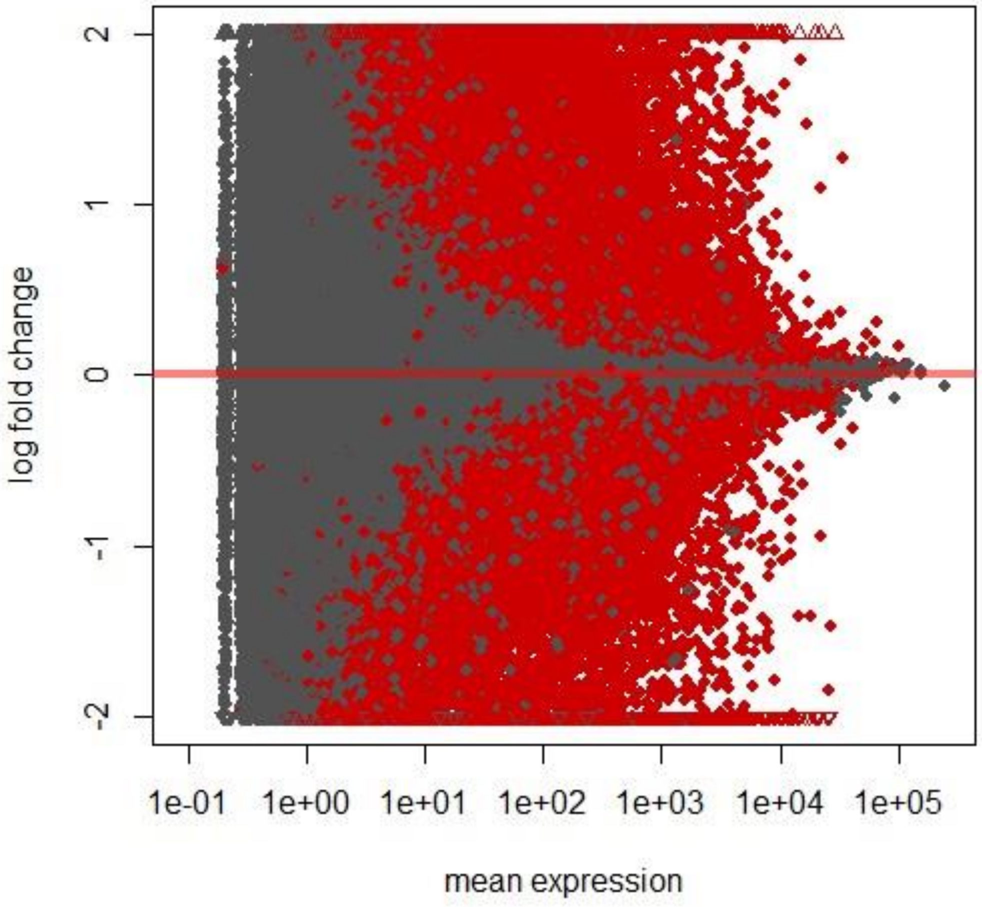

Figure 4.4. LFC of differentially expressed exons compared to mean expression for the analysis using D2 aligned D2 samples. Red indicates significantly differentially expressed genes.

Gene Ontology – B6 Aligned

B6 Aligned Biological Processes

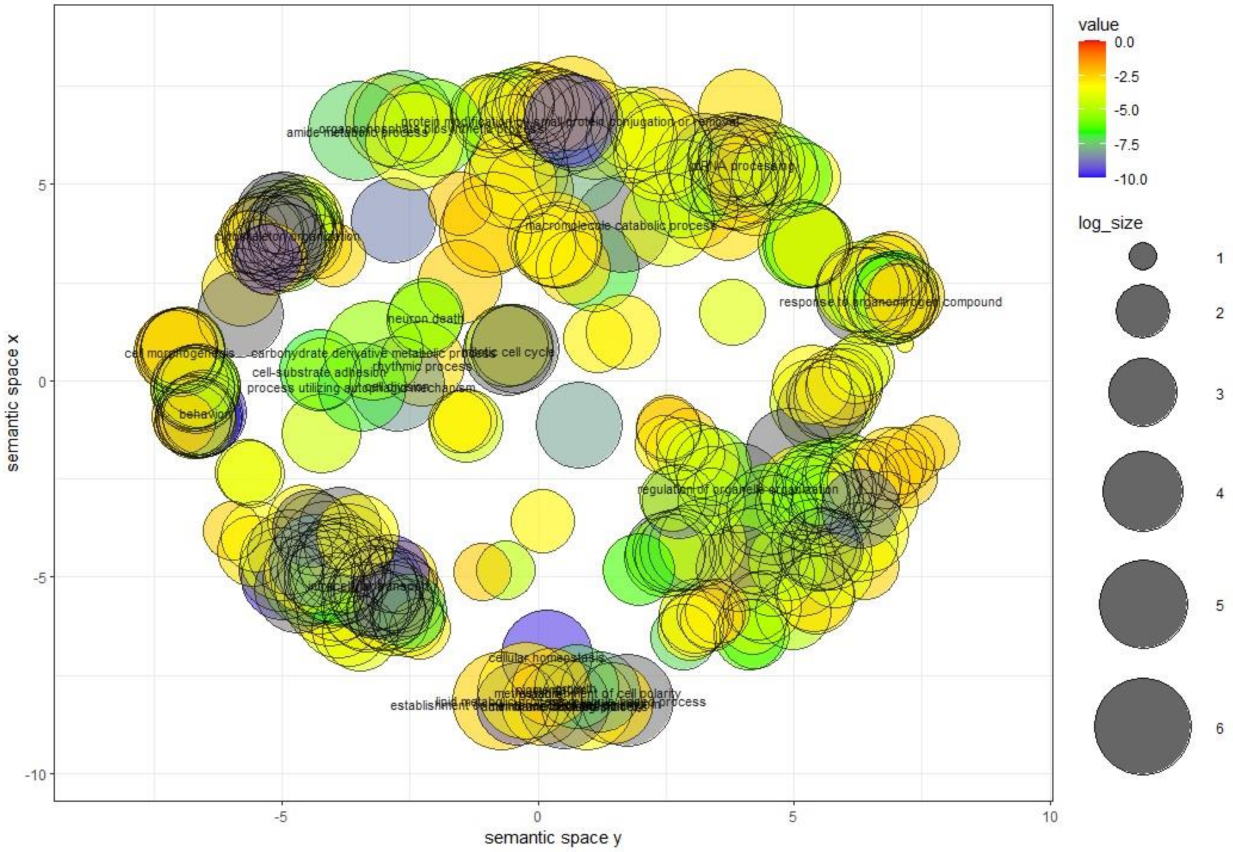

Figure 4.5. Scatterplot of the semantic similarity analysis performed on the biological processes results from the genes with significantly differentially utilized exons from the analysis using B6 aligned D2 samples. Color indicates the log base 10 of the p value output during the ToppFun analysis, with blue indicating the most significantly differentially expressed genes. The size of each point (log_size) indicates the log base 10 of the number of annotations for GO Term ID in selected species in the EBI GOA database.

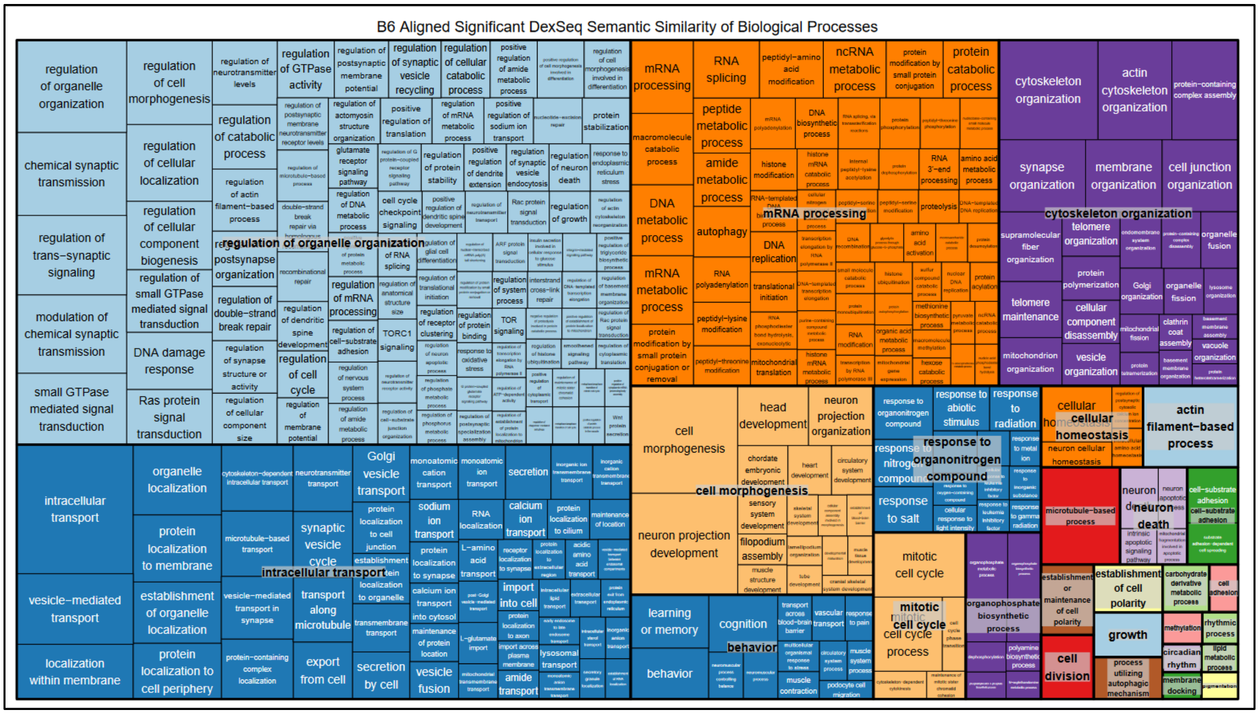

Figure 4.6. Treemap of the semantic similarity analysis performed on the biological processes results from the genes with significantly differentially utilized exons from the analysis using B6 aligned D2 samples. Gene ontology categories are grouped by semantic similarity with closely related categories being clustered together.

B6 Aligned Molecular Functions

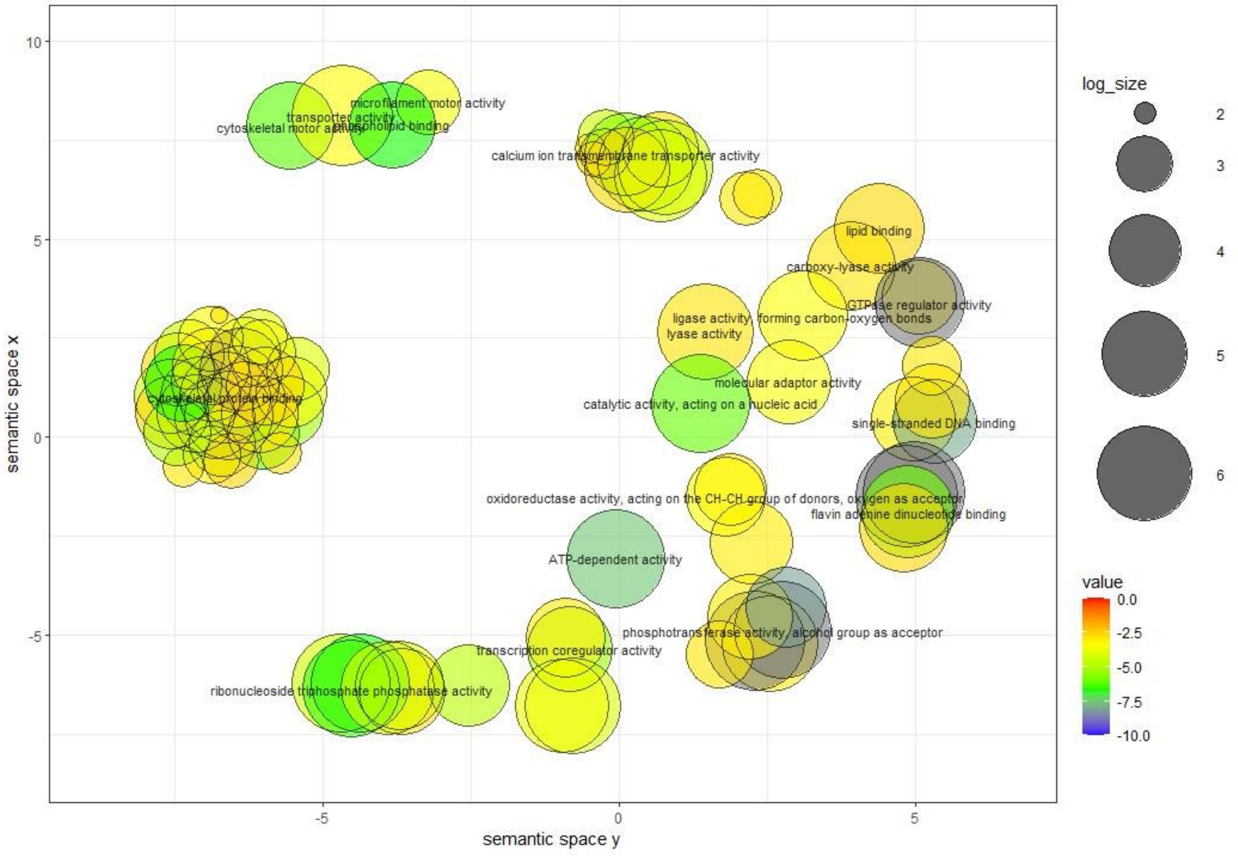

Figure 4.7. Scatterplot of the semantic similarity analysis performed on the molecular function results from the genes with significantly differentially utilized exons from the analysis using B6 aligned D2 samples. Color indicates the log base 10 of the p value output during the ToppFun analysis, with blue indicating the most significantly differentially expressed genes. The size of each point (log_size) indicates the log base 10 of the number of annotations for GO Term ID in selected species in the EBI GOA database.

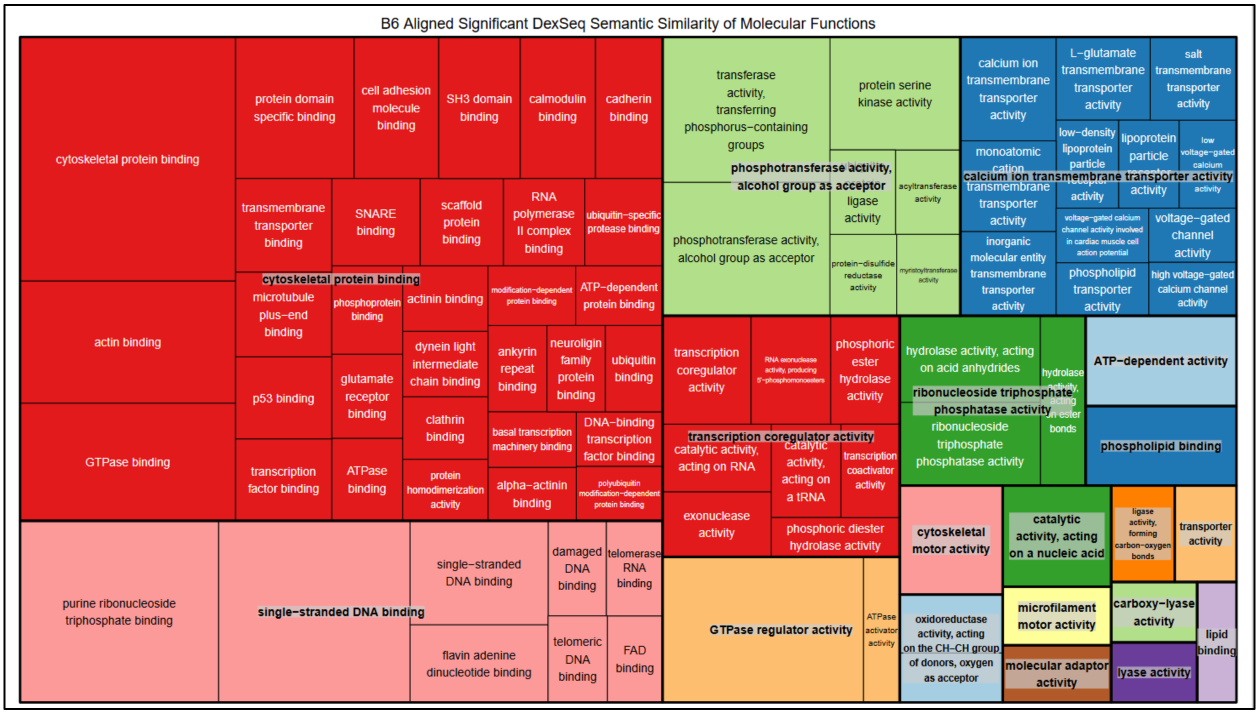

Figure 4.8. Treemap of the semantic similarity analysis performed on the molecular function results from the genes with significantly differentially utilized exons from the analysis using B6 aligned D2 samples. Gene ontology categories are grouped by semantic similarity with closely related categories being clustered together.

B6 Aligned Cellular Components

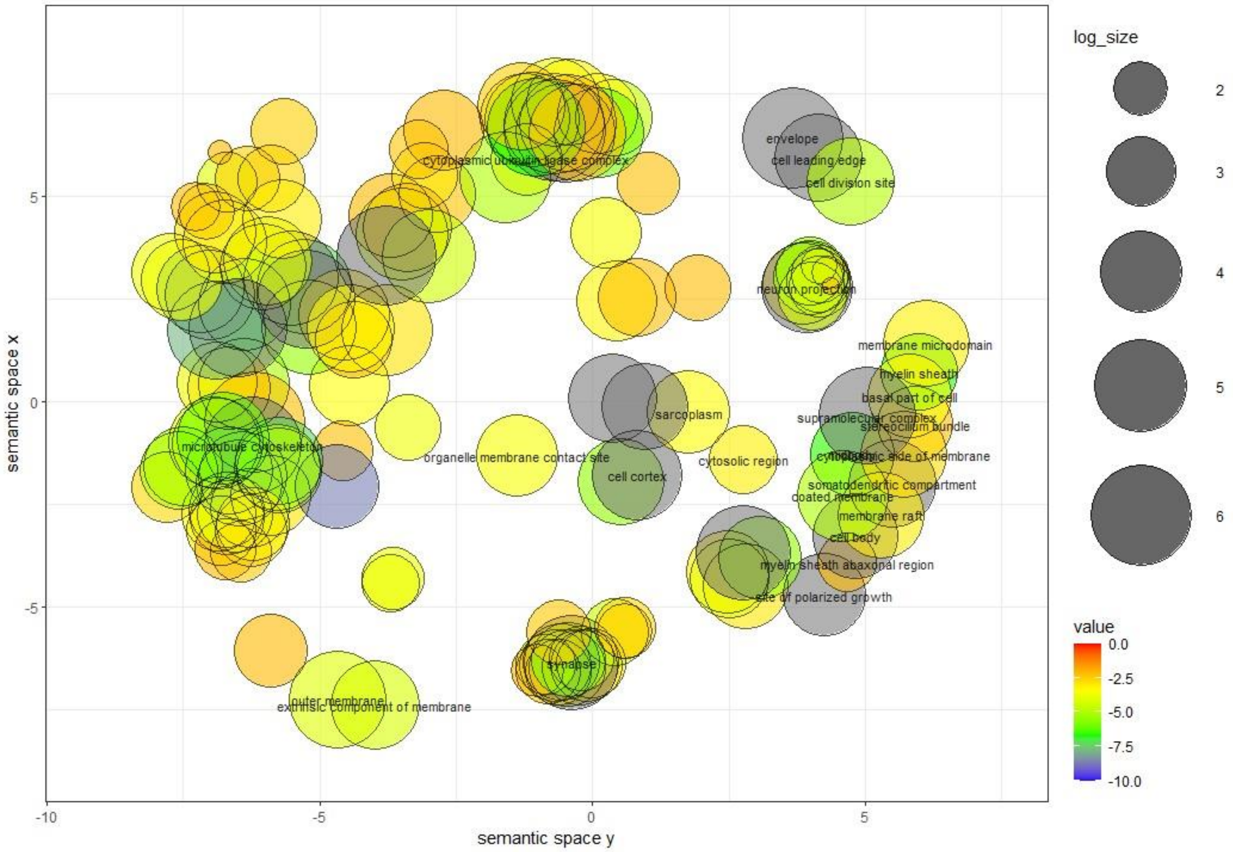

Figure 4.9. Scatterplot of the semantic similarity analysis performed on the cellular component results from the genes with significantly differentially utilized exons from the analysis using B6 aligned D2 samples. Color indicates the log base 10 of the p value output during the ToppFun analysis, with blue indicating the most significantly differentially expressed genes. The size of each point (log_size) indicates the log base 10 of the number of annotations for GO Term ID in selected species in the EBI GOA database.

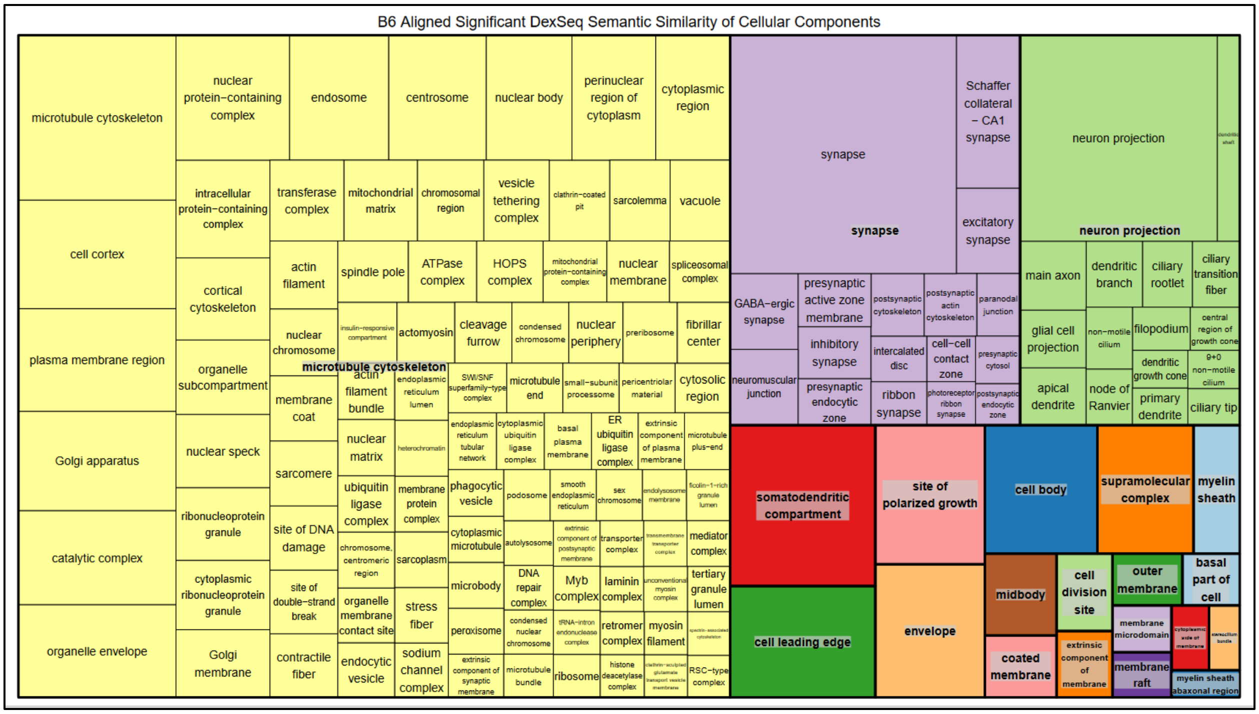

Figure 4.10. Treemap of the semantic similarity analysis performed on the cellular component results from the genes with significantly differentially utilized exons from the analysis using B6 aligned D2 samples. Gene ontology categories are grouped by semantic similarity with closely related categories being clustered together.

Gene Ontology – D2 Aligned

D2 Aligned Biological Processes

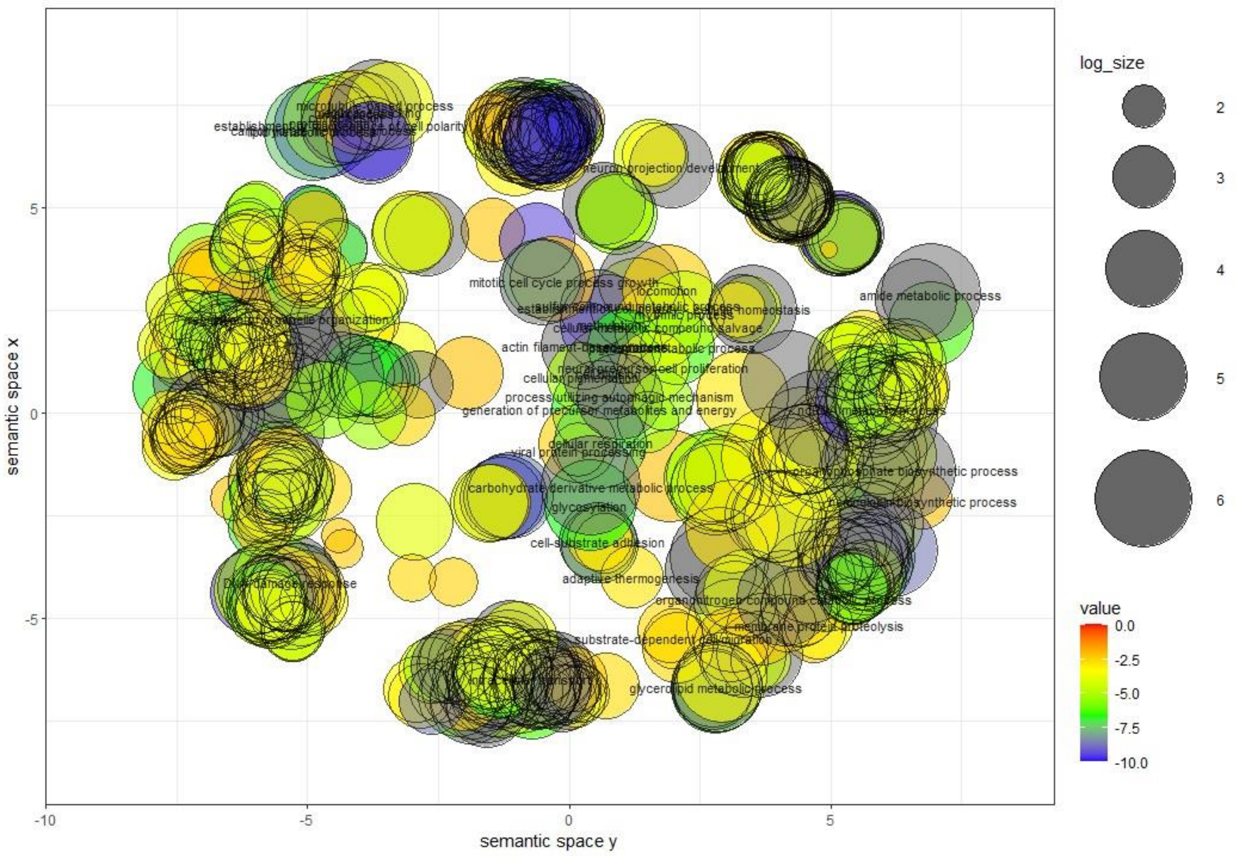

Figure 4.11. Scatterplot of the semantic similarity analysis performed on the biological processes results from the genes with significantly differentially utilized exons from the analysis using D2 aligned D2 samples. Color indicates the log base 10 of the p value output during the ToppFun analysis, with blue indicating the most significantly differentially expressed genes. The size of each point (log_size) indicates the log base 10 of the number of annotations for GO Term ID in selected species in the EBI GOA database.

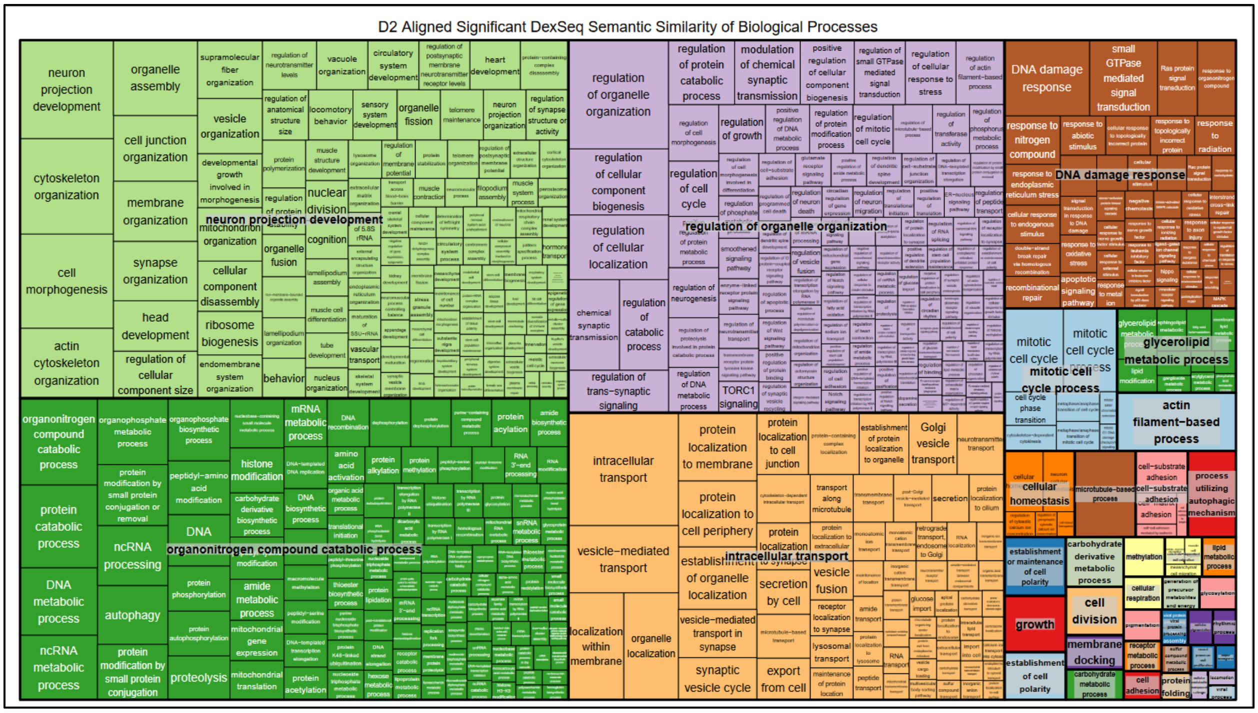

Figure 4.12. Treemap of the semantic similarity analysis performed on the biological processes results from the genes with significantly differentially utilized exons from the analysis using D2 aligned D2 samples. Gene ontology categories are grouped by semantic similarity with closely related categories being clustered together.

D2 Aligned Molecular Functions

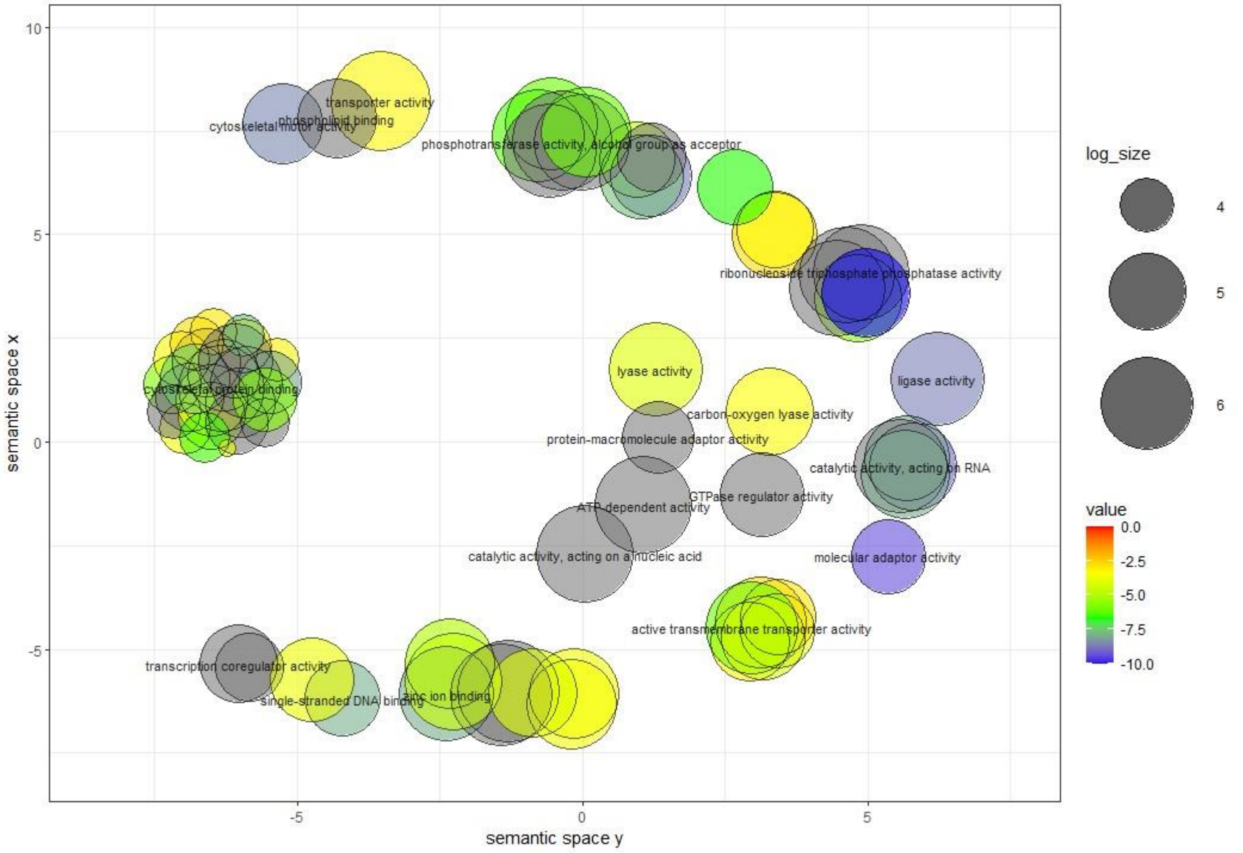

Figure 4.13. Scatterplot of the semantic similarity analysis performed on the molecular function results from the genes with significantly differentially utilized exons from the analysis using D2 aligned D2 samples. Color indicates the log base 10 of the p value output during the ToppFun analysis, with blue indicating the most significantly differentially expressed genes. The size of each point (log_size) indicates the log base 10 of the number of annotations for GO Term ID in selected species in the EBI GOA database.

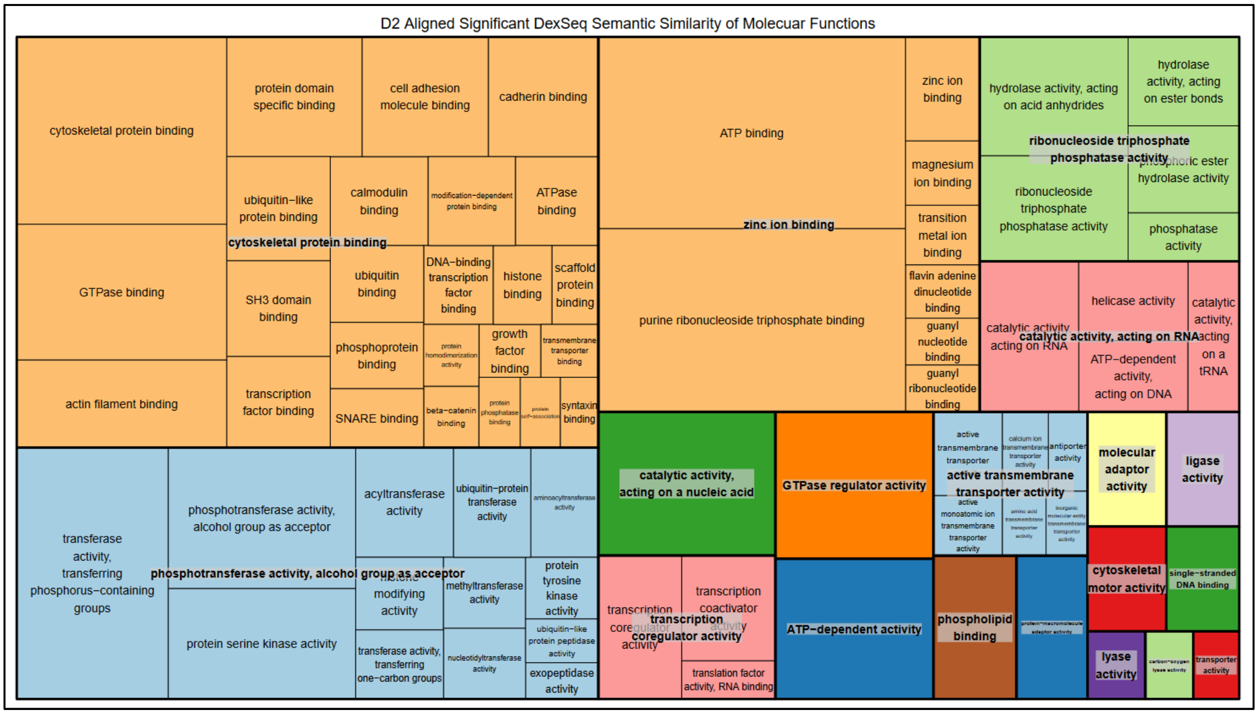

Figure 4.14. Treemap of the semantic similarity analysis performed on the molecular function results from the genes with significantly differentially utilized exons from the analysis using D2 aligned D2 samples. Gene ontology categories are grouped by semantic similarity with closely related categories being clustered together.

D2 Aligned Cellular Components

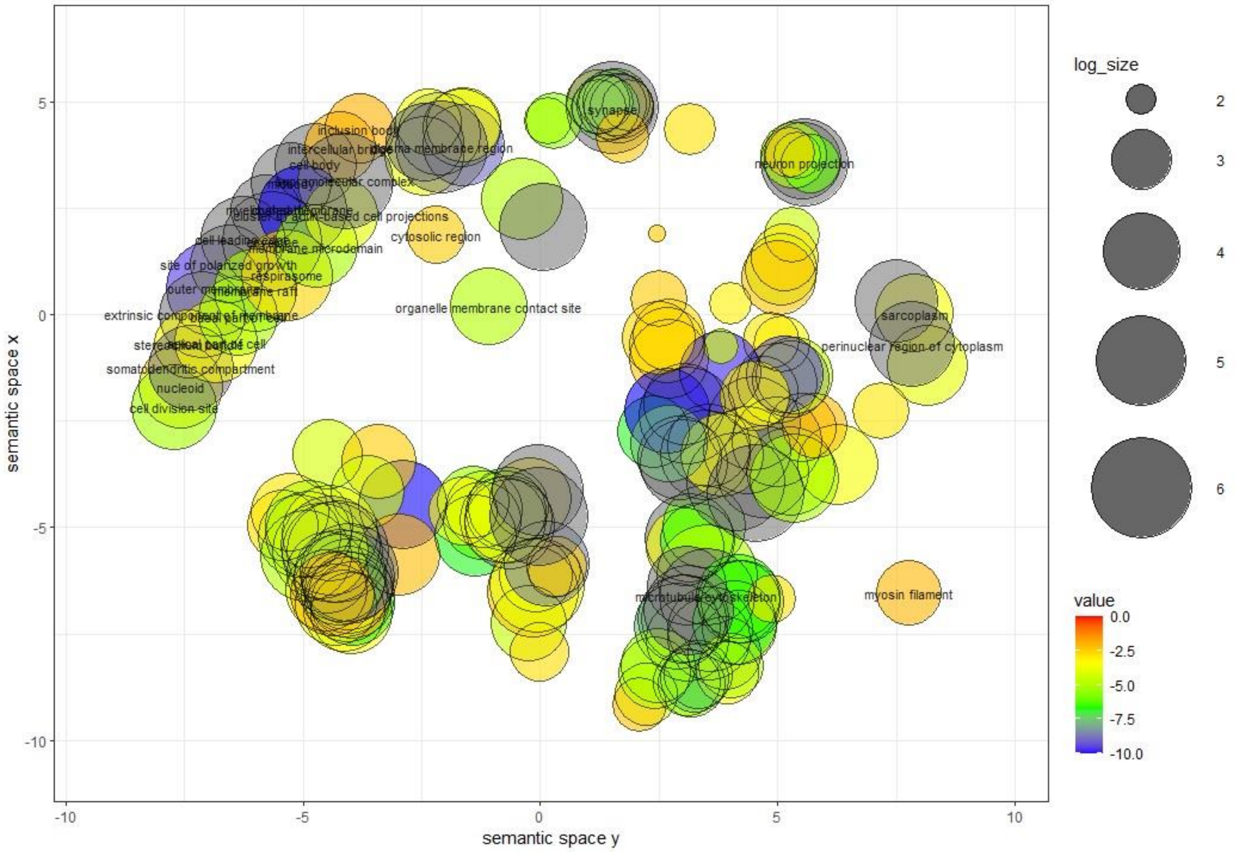

Figure 4.15. Scatterplot of the semantic similarity analysis performed on the cellular component results from the genes with significantly differentially utilized exons from the analysis using D2 aligned D2 samples. Color indicates the log base 10 of the p value output during the ToppFun analysis, with blue indicating the most significantly differentially expressed genes. The size of each point (log_size) indicates the log base 10 of the number of annotations for GO Term ID in selected species in the EBI GOA database.

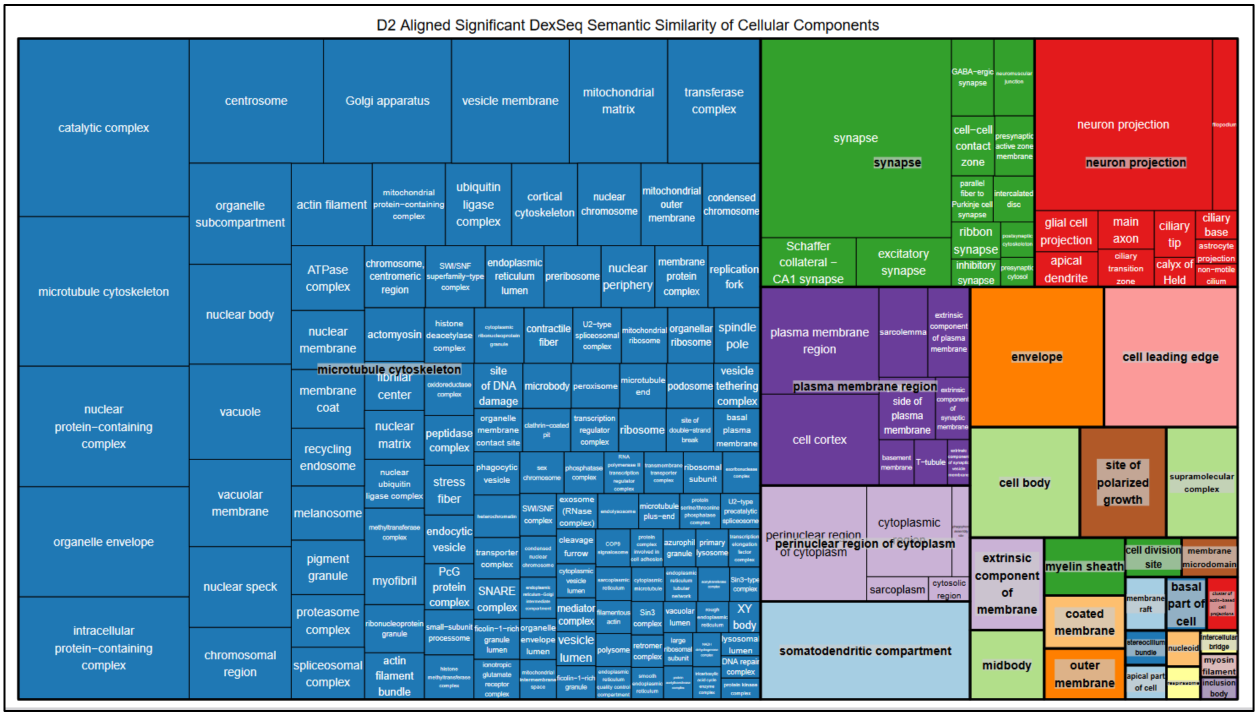

Figure 4.16. Treemap of the semantic similarity analysis performed on the cellular component results from the genes with significantly differentially utilized exons from the analysis using D2 aligned D2 samples. Gene ontology categories are grouped by semantic similarity with closely related categories being clustered together.

Comparison of Results

The first comparison is the number of exons found to have differential utilization, and the number of genes connected to those exons. In the B6 aligned analysis, there were 21,223 differentially utilized exons. Of those, 14,245 (67.12%) were also differentially utilized in the D2 analysis, with 6,978 (32.88%) being unique to the B6 aligned DexSeq analysis. Of the 81,206 differentially utilized exons identified in the D2 aligned DexSeq, 66,961 (82.46%) were unique to the D2 aligned analysis.

The B6 aligned analysis showed 6,650 genes with differentially utilized exons. 5,828 (87.64%) of those were also identified in the D2 aligned analysis, with 822 (12.36%) being unique to the B6 aligned analysis. The D2 aligned analysis had 7,693 (56.90%) unique genes identified only in the D2 aligned analysis.

The magnitudes of the LFCs were broken into positive and negative groups in order to compare them using a t-test. The average positive LFC was had a significantly larger magnitude in the D2 aligned analysis ($p < 0.0001$), and the average negative LFC also had a significantly larger magnitude in the D2 aligned analysis ($p = 0.0021$).

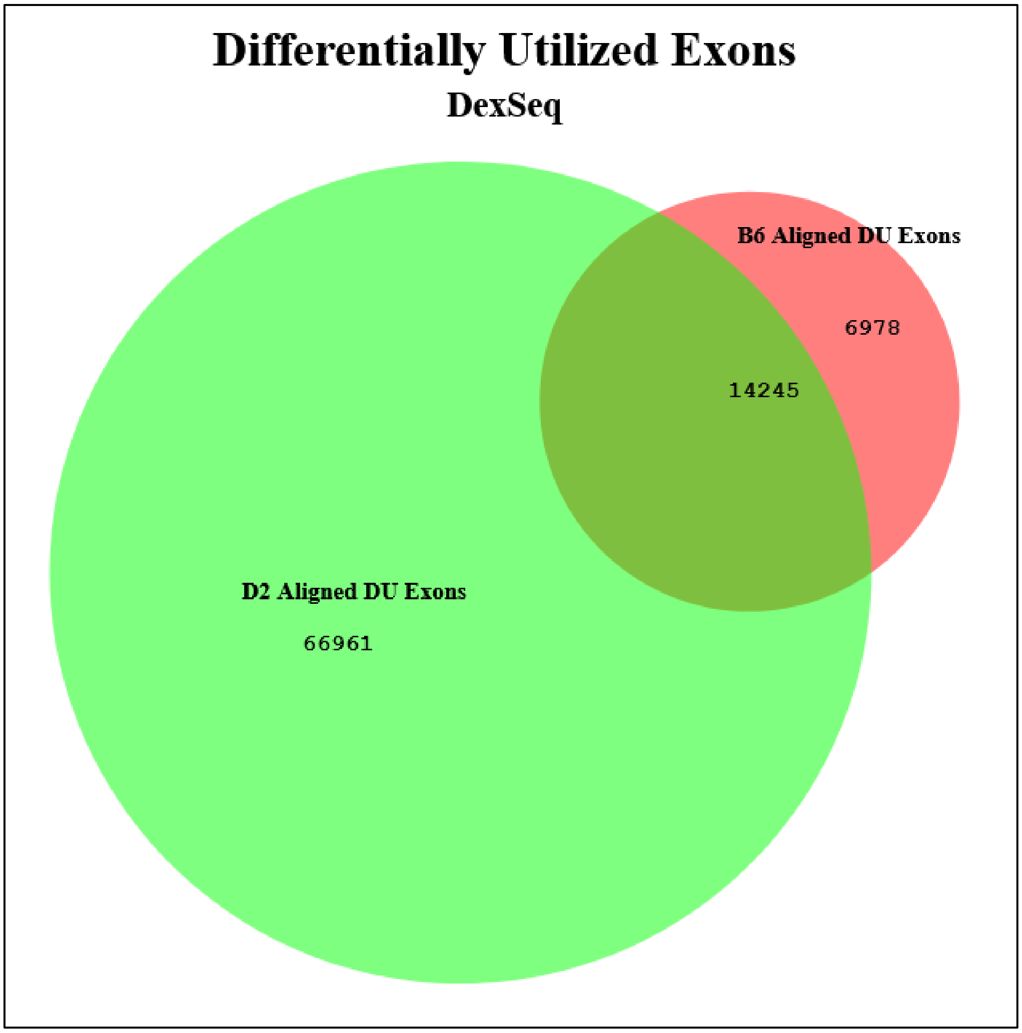

Figure 4.17. Comparison of significantly differentially utilized exons between the B6 aligned D2 analysis (red) and the D2 aligned D2 analysis (green).

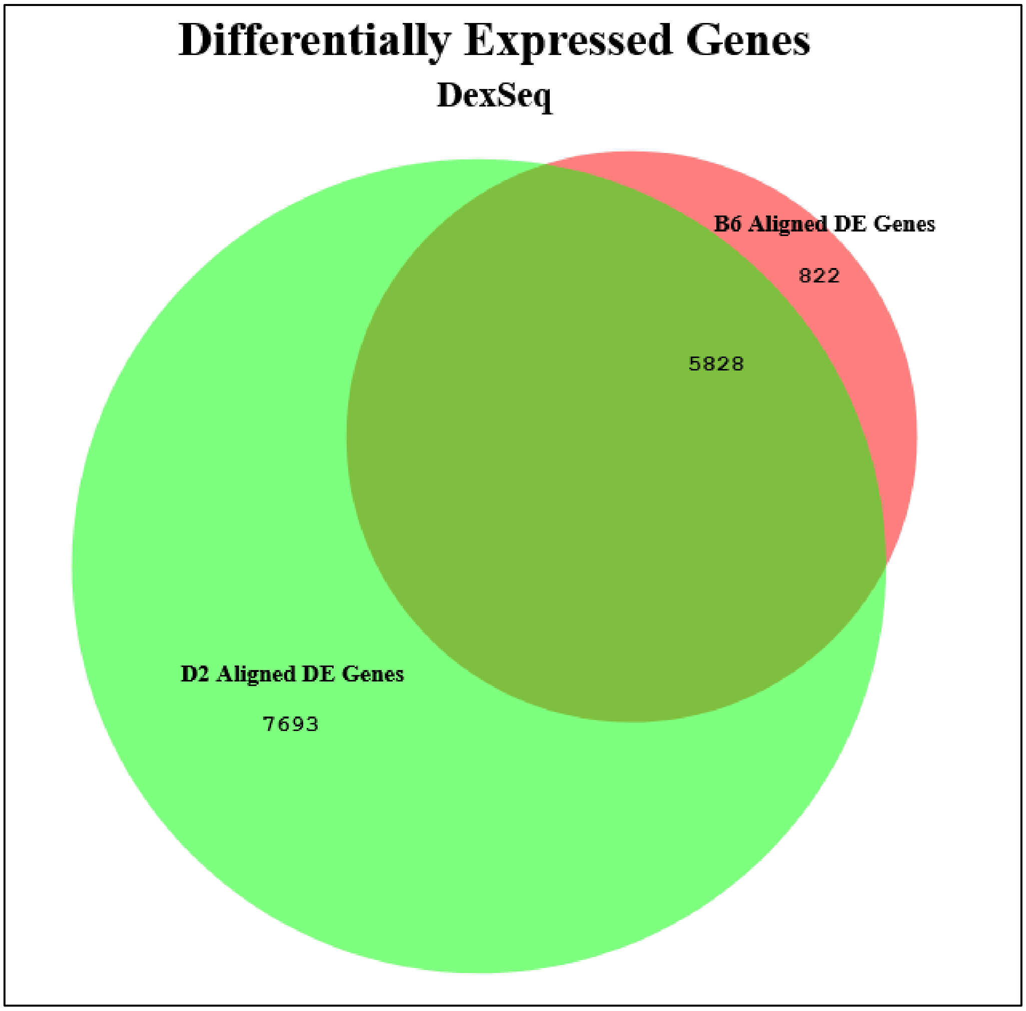

Figure 4.18. Differentially expressed genes identified during the DexSeq analysis. This represents genes only, not exons. Green represents the D2 aligned D2 analysis, and red the B6 aligned D2.

Comparison of Gene Ontology

The semantic similarity analysis run using Revigo shows that there is a large amount overlap in the clustered categories between the B6 aligned and D2 aligned analyses. An analysis of the individual gene ontology terms was used to quantify this, with 49% or more terms overlapping in each category (Table 4.1).

Table 4.1. Number of gene ontology terms in each category for each analysis (B6 aligned and D2 aligned). Overlap is the number of terms found in both sets of results, with percentages for each.

	DexSeq Gene Ontology Terms and Overlap				
	B6 Aligned	D2 Aligned	Overlap	% Overlap (B6)	% Overlap (D2)
Biological Processes	973	1665	829	85.20	49.79
Molecular Function	134	140	82	61.19	58.57
Cellular Component	285	380	243	85.26	63.95

Comparison of Specific Genes

The three specific genes chosen as examples show three different effects from the D2 aligned analysis. Ninein is a gene that has been shown to have differential exon utilization by other studies done in the Miles laboratory. The D2 aligned analysis (Figure 4.20) identified two of the differentially utilized exons (34, 41) that were identified in the B6 aligned analysis (Figure 4.19) with one exon being unique to the B6 aligned analysis (16). However, exon 41, while remaining significant, went from showing higher utilization in B6 in the B6 aligned analysis to showing higher utilization D2 in the D2 aligned analysis. The D2 aligned analysis also identified several unique exons (10, 30, 33, 37, 43, 53) that were not identified in the B6 aligned analysis.

Gabra2 is a gene that has a known differential splicing event between the B6 and D2 strains (Cite, Add specific exon). The D2 aligned analysis (Figure 4.22.) showed the same differentially utilized exons (4, 11, 12, 13, 14, 16, 17, 20, 21) as the B6 aligned analysis (Figure 4.21.) with one exception that was found only in the B6 aligned analysis (10). Two unique exons found only in the D2 aligned analysis (7, 8).

Gsk3b was not found to have differential exon utilization in the B6 aligned analysis, but was found to have 5 differentially utilized exons (3, 9, 10, 11, 14) in the D2 aligned analysis (Figure 4.23).

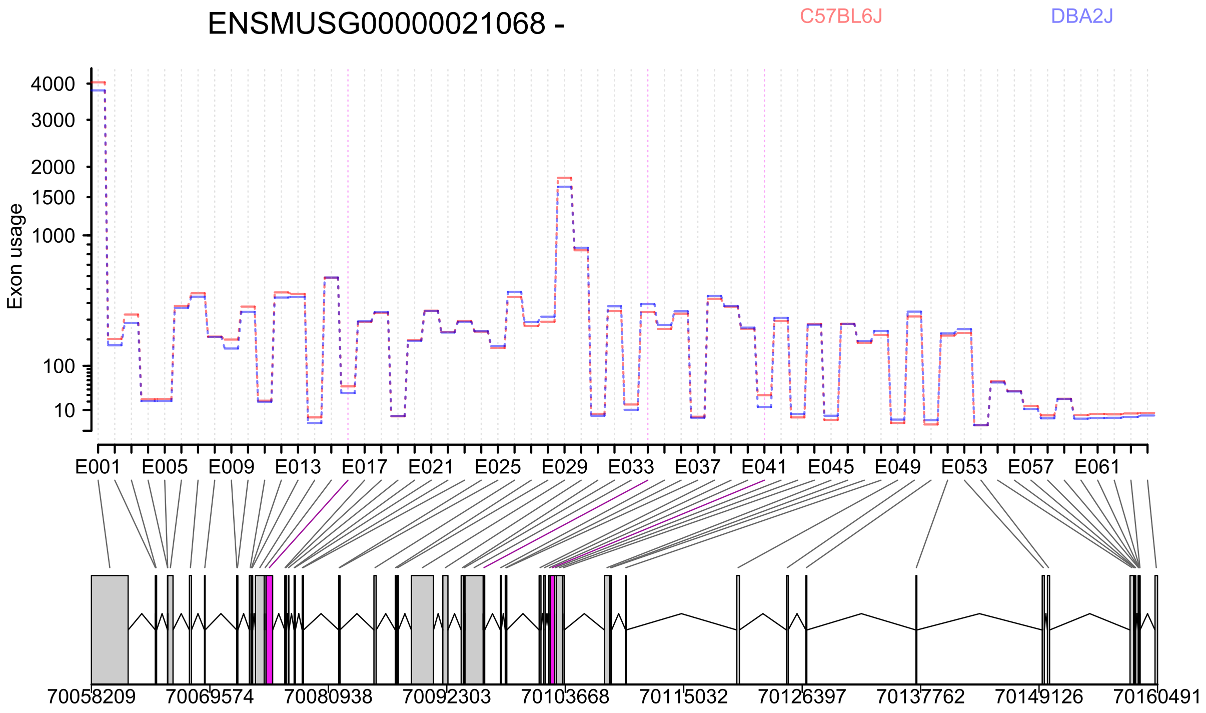

Figure 4.19. B6 aligned DEXSeq splicing event analysis for Ninein. Exons 16 and 41 showed low utilization in both strains, and were determined to be retained introns using a BLAST search. Exon 33 was an alternative splicing even, showing higher utilization in both strains.

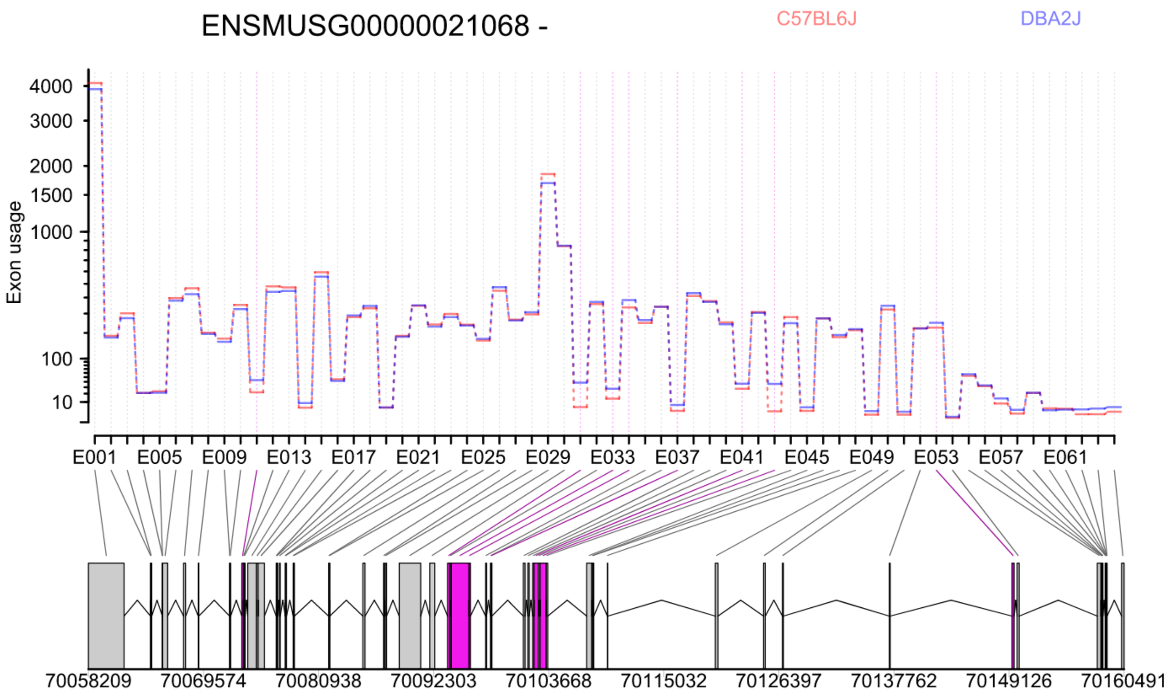

Figure 4.20. D2 aligned DEXSeq splicing event analysis for Ninein. Exon 16 no longer shows differential utilization, and there are multiple new significant events not shown in the B6 aligned analysis. Exons 41 and 34 both show differential utilization, however exon 41 shows lower utilization in B6 in the D2 aligned analysis compared to showing lower utilization in D2 in the B6 aligned analysis.

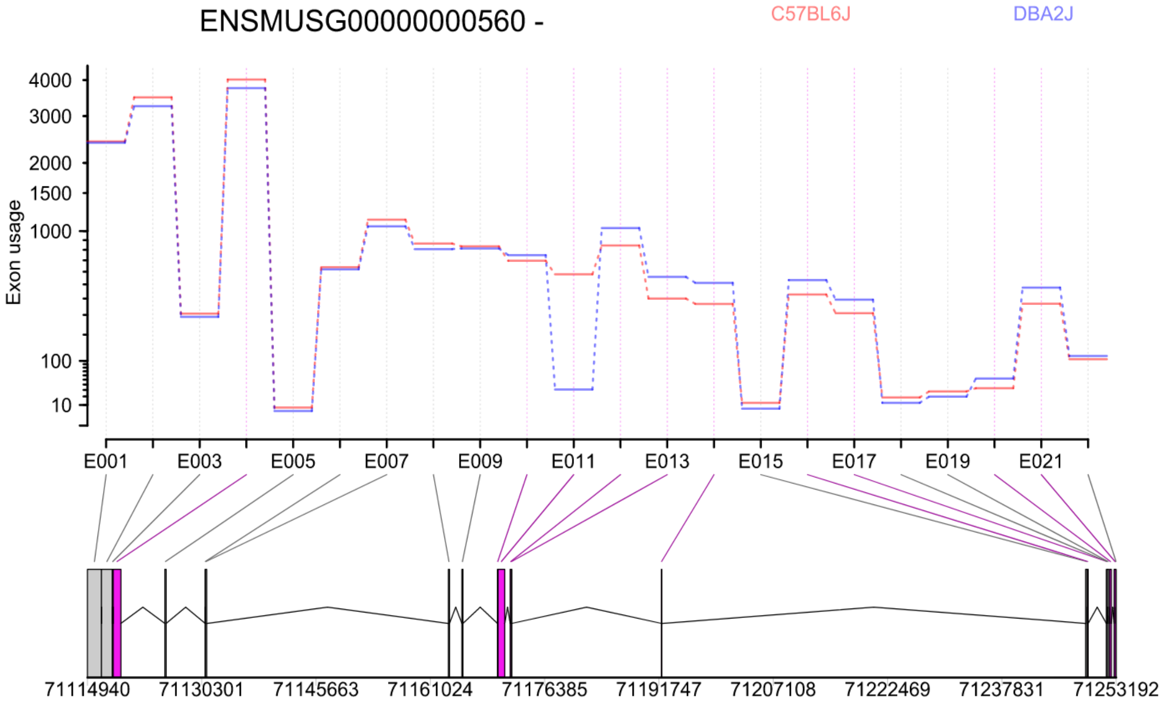

Figure 4.21. B6 aligned DEXSeq splicing event analysis for Gabra2. Gabra2 is has a well know differential splicing event between B6 and D2, a single deleted base pair in an intron, located between exon 3 and 4.

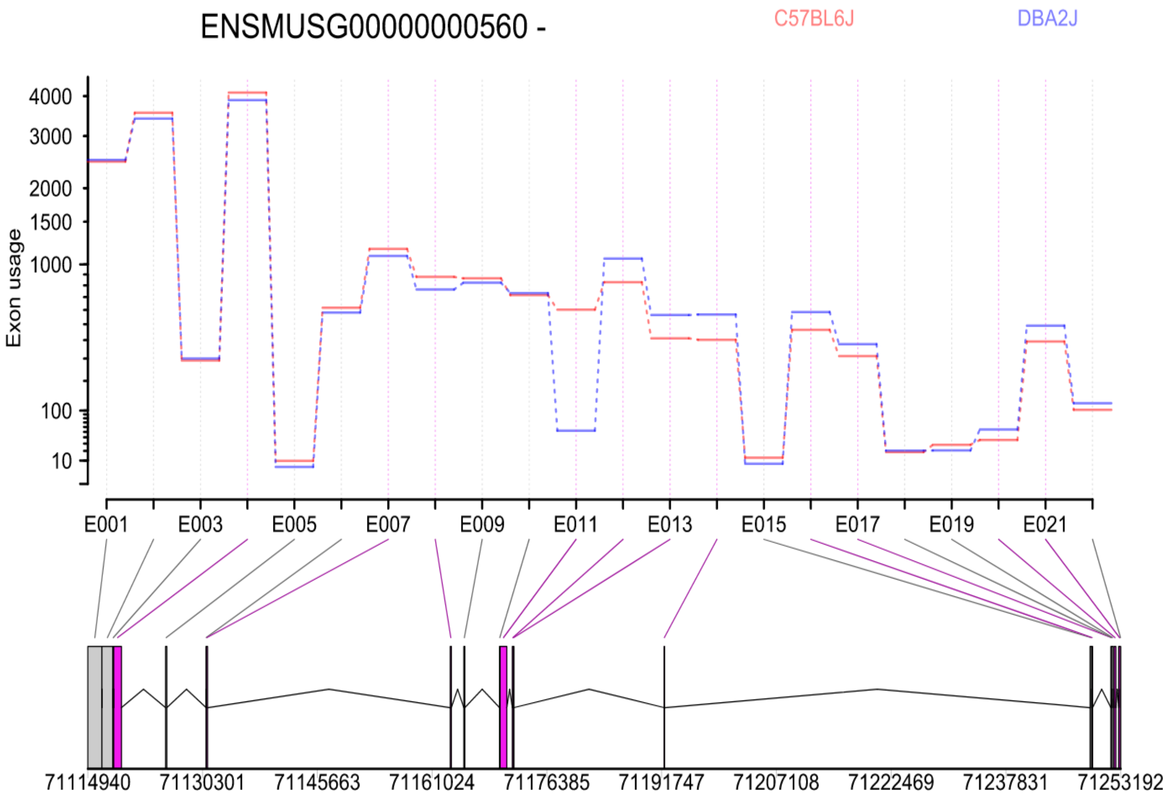

Figure 4.22. D2 aligned DEXSeq splicing event analysis for Gabra2. Gabra2 is has a well know differential splicing event between B6 and D2, located between exons 3 and 4. It can be seen that the B6 and D2 aligned DexSeq analyses identified the same differentially utilized exons, with the D2 aligned analysis identifying slightly more events.

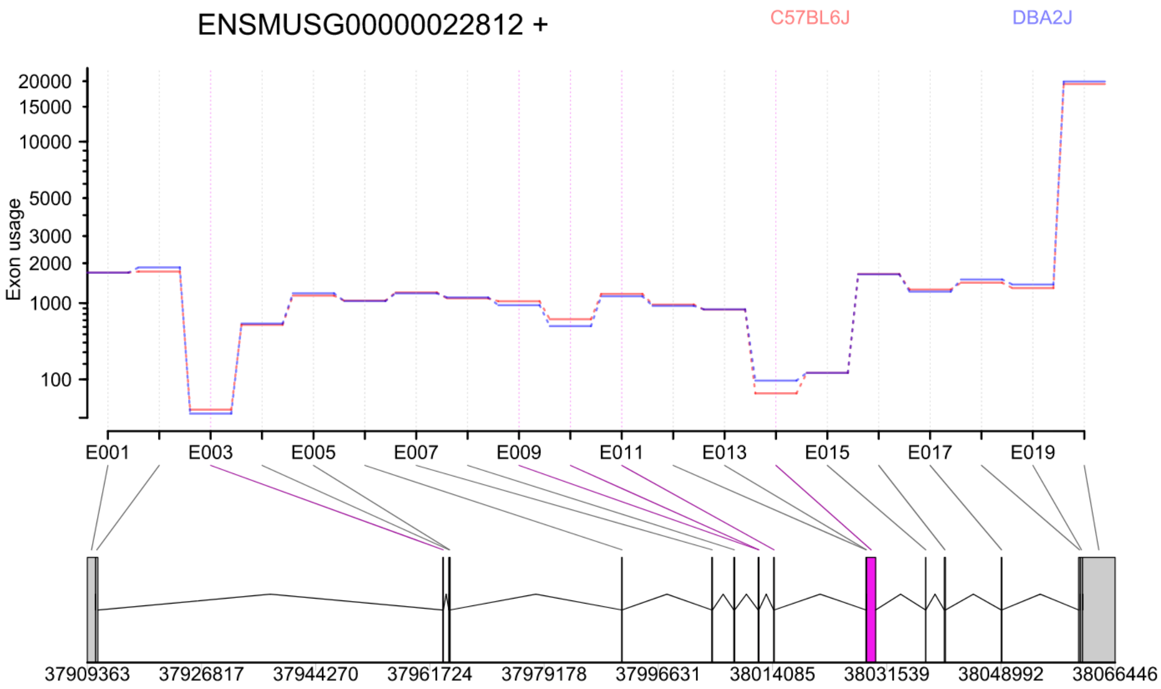

Figure 4.23. D2 aligned DexSeq analysis of Gsk3b. Gsk3b was not shown to have differential exon utilization in the B6 aligned analysis, but the D2 aligned analysis identified several exons that were differentially utilized.

Discussion

Count Data Preparation

With the aid of Dr. Dozmorov, the DexSeq counts were generated using the D2 aligned annotation provided by Dr. Keane. However, that annotation would not work for the DexSeq analysis, so the B6 annotation was used. Because the count data was generated using D2 aligned D2, and the resulting D2 Ensembl IDs were converted to their B6 counterparts, using the B6 annotation for the analysis is acceptable. In order to verify the validity of this step, a future analysis using the D2 annotation with the B6 ensemble IDs converted to their D2 counterparts should be run. The results of that should be very similar, though some variance is to be expected when using a different annotation. For now, this is an acceptable method of using the D2 aligned counts in DexSeq.

DexSeq Results – B6 Aligned and D2 Aligned

In both the B6 and D2 aligned analyses, the LFCs were evenly distributed between positive and negative, with the D2 aligned analysis skewing slightly towards the positive. This indicates that the analysis did not have significant bias towards either positive or negative LFC that would affect the results or indicate an error in the analysis.

In both the B6 (Figure 4.1) and the D2 (Figure 4.3) aligned analysis, there is high correlation between out exonic regions, with B6 and D2 correlating more to themselves than to each other, though the correlation between the two is still quite high. There is low correlation between the out exonic regions and the rest of the exons, though the “rest of the exons” correlate strongly to each other. Interestingly one of the B6 samples correlates more strongly to the D2 than to the B6, though not enough to be an outlier. This is to be expected, as the samples are taken from the same species, and this indicates that it is possible to differentiate between closely related substrains. The out exonic regions having low correlation with the rest of the exons also indicates that the analysis was done correctly.

Comparison of Results

Similarly to the differential expression analysis, the D2 aligned analysis showed significantly larger numbers of differentially utilized exons, and a correspondingly larger number of genes with significantly differentially utilized exons. The retention among genes was as good as the differential expression analysis, with 87.64% being identified in both the B6 and D2 aligned analyses. The exon overlap was much lower, at 67.12%, but with a much larger number of unique differentially utilized exons. This indicates that the D2 aligned analysis does provide an improvement in the identification of differentially utilized exons, though the lack of retention from the B6 aligned analysis warrants further investigation. It may be related to the usage of the B6 annotation with the D2 aligned counts, and this will be tested in the future.

The magnitude of the LFCs was higher in the D2 aligned analysis for both positive and negative LFCs. This indicates that not only were more differentially utilized exons identified in the D2 aligned analysis, those exons were also significantly more expressed or less expressed in the D2 aligned analysis. This indicates potential improvement, as the analysis run using D2 aligned counts showed a significant difference from the B6 aligned counts. It is important to quantify which analysis is “better”, however. Future analysis will look more into differential exons and gene expression, specifically focusing on splicing events to determine this. Emma Gnatowski has begun this analysis already. Another goal is to look more deeply into the exons identified in both analyses, with a focus on the unique exons. If they follow a similar pattern to the overlapping exons in size and LFC, then that removes a potential factor causing the D2 aligned analysis to identify them. Eventually all factors other than the D2 alignment will be accounted for and a definitive answer will be found.

Comparison of Gene Ontology

The high amount of overlap is what was expected and is encouraging to see. Because the D2 aligned analysis had much larger numbers of significant GO terms (Table 4.1.) while still having high overlap with the B6 aligned analysis, it can be inferred that aligning to the D2 provides a noticeable change in the results while not losing results found in the B6 aligned analysis.

Comparison of Specific Genes

The specific genes compared are either currently being studied in the Miles laboratory or, in the case of Gabra2, have a known alternative splicing event.

With Ninein, the most interesting result is the flip of exon 41. In the B6 aligned analysis, it showed higher utilization in the B6 samples, whereas in the D2 aligned analysis it showed higher utilization in the D2 samples. This warrants further investigation, as this indicates not only a significant change in the magnitude of the event, but a complete reversal in the direction. The D2 aligned analysis also identified unique differentially utilized exons, and these should be tested as described above to determine if there are any other factors causing these to be missed in the B6 aligned analysis.

Gabra2 showed the same directionality of each change, and the D2 aligned analysis identified several new differentially utilized exons. This is in keeping with the results seen before, and it is good to see that the differentially utilized exons from the B6 aligned analysis, including a known deletion that occurred and became fixed in the B6 line, located in the intro between exons 3 and 4 (Mulligan et al., 2019). This shows that the D2 aligned analysis is successfully identifying differentially utilized exons, and not simply giving false positives.

With Gsk3b, it was not found to have differential exon utilization in the B6 aligned analysis. The D2 aligned analysis did however find several exons with differential utilization. This shows why aligning D2 mice to the D2 genome for these analyses is important, as it can identify differentially expressed exons and genes that would have otherwise been missed.

Bibliography

Agarwalla, S., Arroyo, N. S., Long, N. E., O'Brien, W. T., Abel, T., & Bandyopadhyay, S. (2020). Male-specific alterations in structure of isolation call sequences of mouse pups with 16p11.2 deletion. *Genes, Brain and Behavior*, *19*(7), e12681. https://doi.org/10.1111/gbb.12681

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

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

Anjum, A., Jaggi, S., Varghese, E., Lall, S., Bhowmik, A., & Rai, A. (2016). Identification of Differentially Expressed Genes in RNA-seq Data of *Arabidopsis thaliana*: A Compound Distribution Approach. *Journal of Computational Biology*, *23*(4), 239–247. https://doi.org/10.1089/cmb.2015.0205

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. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Bahi, A., & Dreyer, J.-L. (2012). Involvement of nucleus accumbens dopamine D1 receptors in ethanol drinking, ethanol-induced conditioned place preference, and ethanol-induced psychomotor sensitization in mice. *Psychopharmacology*, *222*(1), 141–153. https://doi.org/10.1007/s00213-011-2630-8

Blake, J. A., Baldarelli, R., Kadin, J. A., Richardson, J. E., Smith, C. L., Bult, C. J., & Mouse Genome Database Group. (2021). Mouse Genome Database (MGD): Knowledgebase for mouse-human comparative biology. *Nucleic Acids Research*, *49*(D1), D981–D987. https://doi.org/10.1093/nar/gkaa1083

Bottomly, D., Walter, N. A. R., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., Searles, R. P., Mooney, M., McWeeney, S. K., & Hitzemann, R. (2011). Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays. *PLoS ONE*, *6*(3), e17820. https://doi.org/10.1371/journal.pone.0017820

Chen, J., Bardes, E. E., Aronow, B. J., & Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, *37*(suppl_2), W305–W311. https://doi.org/10.1093/nar/gkp427

Choi, K., He, H., Gatti, D. M., Philip, V. M., Raghupathy, N., Gyuricza, I. G., Munger, S. C., Chesler, E. J., & Churchill, G. A. (2020). *Genotype-free individual genome reconstruction of Multiparental Population Models by RNA sequencing data* [Preprint]. Bioinformatics. https://doi.org/10.1101/2020.10.11.335323

Cotto, K. C., Feng, Y.-Y., Ramu, A., Richters, M., Freshour, S. L., Skidmore, Z. L., Xia, H., McMichael, J. F., Kunisaki, J., Campbell, K. M., Chen, T. H.-P., Rozycki, E. B., Adkins, D., Devarakonda, S., Sankararaman, S., Lin, Y., Chapman, W. C., Maher, C. A., Arora, V., … Griffith, M. (2018). *RegTools: Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer* [Preprint]. Bioinformatics. https://doi.org/10.1101/436634

Dainat, J., Hereñú, D., Dr. K. D. Murray, Davis, E., Crouch, K., LucileSol, Agostinho, N., Pascal-Git, Zollman, Z., & Tayyrov. (2023). *NBISweden/AGAT: AGAT-v1.2.0* (v1.2.0) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.3552717

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Daniels, G. M., & Buck, K. J. (2002). Expression profiling identifies strain-specific changes associated with ethanol withdrawal in mice. *Genes, Brain, and Behavior*, *1*(1), 35–45. https://doi.org/10.1046/j.1601-1848.2001.00008.x

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

Doran, A. G., Wong, K., Flint, J., Adams, D. J., Hunter, K. W., & Keane, T. M. (2016). Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation and homozygous truncating mutations. *Genome Biology*, *17*(1), 167. https://doi.org/10.1186/s13059-016-1024-y

Erickson, A., Zhou, S., Luo, J., Li, L., Huang, X., Even, Z., Huang, H., Xu, H.-M., Peng, J., Lu, L., & Wang, X. (2021). Genetic architecture of protein expression and its regulation in the mouse brain. *BMC Genomics*, *22*(1), 875. https://doi.org/10.1186/s12864-021-08168-y

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics*, *8*(4), 286–298. https://doi.org/10.1038/nrg2005

García-García, M. J. (2020). A History of Mouse Genetics: From Fancy Mice to Mutations in Every Gene. *Advances in Experimental Medicine and Biology*, *1236*, 1–38. https://doi.org/10.1007/978-981-15-2389-2_1

Gremel, C. M., & Cunningham, C. L. (2008). Roles of the Nucleus Accumbens and Amygdala in the Acquisition and Expression of Ethanol-Conditioned Behavior in Mice. *Journal of Neuroscience*, *28*(5), 1076–1084. https://doi.org/10.1523/JNEUROSCI.4520-07.2008

Han, B., Jones, C. M., Einstein, E. B., Powell, P. A., & Compton, W. M. (2021). Use of Medications for Alcohol Use Disorder in the US: Results From the 2019 National Survey on Drug Use and Health. *JAMA Psychiatry*, *78*(8), 922–924. https://doi.org/10.1001/jamapsychiatry.2021.1271

Hoffman, P., & Tabakoff, B. (2005). Gene expression in animals with different acute responses to ethanol. *Addiction Biology*, *10*(1), 63–69. https://doi.org/10.1080/13556210412331308985

Kerns, R. T., Ravindranathan, A., Hassan, S., Cage, M. P., York, T., Sikela, J. M., Williams, R. W., & Miles, M. F. (2005a). Ethanol-Responsive Brain Region Expression Networks: Implications for Behavioral Responses to Acute Ethanol in DBA/2J versus C57BL/6J Mice. *Journal of Neuroscience*, *25*(9), 2255–2266. https://doi.org/10.1523/JNEUROSCI.4372-04.2005

Kerns, R. T., Ravindranathan, A., Hassan, S., Cage, M. P., York, T., Sikela, J. M., Williams, R. W., & Miles, M. F. (2005b). Ethanol-responsive brain region expression networks: Implications for behavioral responses to acute ethanol in DBA/2J versus C57BL/6J mice. *The Journal* *of Neuroscience: The Official Journal of the Society for Neuroscience*, *25*(9), 2255–2266. https://doi.org/10.1523/JNEUROSCI.4372-04.2005

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

Kranzler, H. R., Zhou, H., Kember, R. L., Vickers Smith, R., Justice, A. C., Damrauer, S., Tsao, P. S., Klarin, D., Baras, A., Reid, J., Overton, J., Rader, D. J., Cheng, Z., Tate, J. P., Becker, W. C., Concato, J., Xu, K., Polimanti, R., Zhao, H., & Gelernter, J. (2019). Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nature Communications*, *10*(1), 1499. https://doi.org/10.1038/s41467-019-09480-8

Lê, A. D., Ko, J., Chow, S., & Quan, B. (1994). Alcohol consumption by C57BL/6, BALB/c, and DBA/2 mice in a limited access paradigm. *Pharmacology Biochemistry and Behavior*, *47*(2), 375–378. https://doi.org/10.1016/0091-3057(94)90026-4

Li, Y., Rao, X., Mattox, W. W., Amos, C. I., & Liu, B. (2015). RNA-Seq Analysis of Differential Splice Junction Usage and Intron Retentions by DEXSeq. *PLOS ONE*, *10*(9), e0136653. https://doi.org/10.1371/journal.pone.0136653

Li, Y., & Xie, X. (2013). A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. *BMC Bioinformatics*, *14 Suppl 5*(Suppl 5), S11. https://doi.org/10.1186/1471-2105-14-S5-S11

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

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

Mahnke, C. G., Jänig, U., Werner, J. A., & Rudert, H. (1994). Primary papillary carcinoma of the thyroglossal duct: Case report and review of the literature. *Auris, Nasus, Larynx*, *21*(4), 258–263. https://doi.org/10.1016/s0385-8146(12)80091-5

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics*, *12*(10), 671–682. https://doi.org/10.1038/nrg3068

Mekada, K., Hirose, M., Murakami, A., & Yoshiki, A. (2015). Development of SNP markers for C57BL/6N-derived mouse inbred strains. *Experimental Animals*, *64*(1), 91–100. https://doi.org/10.1538/expanim.14-0061

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. https://doi.org/10.1038/nmeth.1226

Mouse Genome Sequencing Consortium, Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., … Lander, E. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*(6915), 520–562. https://doi.org/10.1038/nature01262

Mulligan, M. K., Abreo, T., Neuner, S. M., Parks, C., Watkins, C. E., Houseal, M. T., Shapaker, T. M., Hook, M., Tan, H., Wang, X., Ingels, J., Peng, J., Lu, L., Kaczorowski, C. C., Bryant, C. D., Homanics, G. E., & Williams, R. W. (2019). Identification of a Functional Non-coding Variant in the GABAA Receptor α2 Subunit of the C57BL/6J Mouse Reference Genome: Major Implications for Neuroscience Research. *Frontiers in Genetics*, *10*, 188. https://doi.org/10.3389/fgene.2019.00188

Mulligan, M. K., Zhao, W., Dickerson, M., Arends, D., Prins, P., Cavigelli, S. A., Terenina, E., Mormede, P., Lu, L., & Jones, B. C. (2018). Genetic Contribution to Initial and Progressive Alcohol Intake Among Recombinant Inbred Strains of Mice. *Frontiers in Genetics*, *9*, 370. https://doi.org/10.3389/fgene.2018.00370

Ng, G. YK., O’Dowd, B. F., & George, S. R. (1994). Genotypic differences in brain dopamine receptor function in the DBA/2J and C57BL/6J inbred mouse strains. *European Journal of Pharmacology: Molecular Pharmacology*, *269*(3), 349–364. https://doi.org/10.1016/0922-4106(94)90043-4

O’Brien, M. A., Weston, R. M., Sheth, N. U., Bradley, S., Bigbee, J., Pandey, A., Williams, R. W., Wolstenholme, J. T., & Miles, M. F. (2018). Ethanol-Induced Behavioral Sensitization Alters the Synaptic Transcriptome and Exon Utilization in DBA/2J Mice. *Frontiers in Genetics*, *9*, 402. https://doi.org/10.3389/fgene.2018.00402

Oshlack, A., Robinson, M. D., & Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biology*, *11*(12), 220. https://doi.org/10.1186/gb-2010-11-12-220

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Piper, M. M., Radhika Khetani, Mary. (2017, May 12). *Gene-level differential expression analysis with DESeq2*. Introduction to DGE - ARCHIVED. https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html

Przybyła, P., Shardlow, M., Aubin, S., Bossy, R., Eckart De Castilho, R., Piperidis, S., McNaught, J., & Ananiadou, S. (2016). Text mining resources for the life sciences. *Database*, *2016*. https://doi.org/10.1093/database/baw145

Ptashne, M., & Gann, A. (1997). Transcriptional activation by recruitment. *Nature*, *386*(6625), 569–577. https://doi.org/10.1038/386569a0

PubMed. (n.d.). PubMed. Retrieved January 30, 2023, from https://pubmed.ncbi.nlm.nih.gov/

Putman, A. H., Wolen, A. R., Harenza, J. L., Yordanova, R. K., Webb, B. T., Chesler, E. J., & Miles, M. F. (2016). Identification of quantitative trait loci and candidate genes for an anxiolytic-like response to ethanol in BXD recombinant inbred strains: Chromosomal loci influencing ethanol anxiolysis. *Genes, Brain and Behavior*, *15*(4), 367–381. https://doi.org/10.1111/gbb.12289

Resnik, P. (2011). *Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language*. https://doi.org/10.48550/ARXIV.1105.5444

Sacks, J. J., Gonzales, K. R., Bouchery, E. E., Tomedi, L. E., & Brewer, R. D. (2015). 2010 National and State Costs of Excessive Alcohol Consumption. *American Journal of Preventive Medicine*, *49*(5), e73–e79. https://doi.org/10.1016/j.amepre.2015.05.031

SAMHSA. (2021). *Table 5.1A – Substance Use Disorder for Specific Substances in Past Year: Among People Aged 12 or Older; by Age Group, Numbers in Thousands, 2021*. https://www.samhsa.gov/data/sites/default/files/reports/rpt39441/NSDUHDetailedTabs2021/NSDUHDetailedTabs2021/NSDUHDetTabsSect5pe2021.htm#tab5.6a

Schlicker, A., Domingues, F. S., Rahnenführer, J., & Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, *7*(1), 302. https://doi.org/10.1186/1471-2105-7-302

Shen, S., Park, J. W., Lu, Z., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., & Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences*, *111*(51). https://doi.org/10.1073/pnas.1419161111

Srivatsa, S., Parthasarathy, S., Molnár, Z., & Tarabykin, V. (2015). Sip1 Downstream Effector ninein Controls Neocortical Axonal Growth, Ipsilateral Branching, and Microtubule Growth and Stability. *Neuron*, *85*(5), 998–1012. https://doi.org/10.1016/j.neuron.2015.01.018

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLOS ONE*, *6*(7), e21800. https://doi.org/10.1371/journal.pone.0021800

Tabakoff, B., & Hoffman, P. L. (2000). Animal models in alcohol research. *Alcohol Research & Health: The Journal of the National Institute on Alcohol Abuse and Alcoholism*, *24*(2), 77–84.

ToppFun—Functional Enrichment. (n.d.). Retrieved January 16, 2023, from https://toppgene.cchmc.org/enrichment.jsp

Wolen, A. R., & Miles, M. F. (2012). Identifying gene networks underlying the neurobiology of ethanol and alcoholism. *Alcohol Research: Current Reviews*, *34*(3), 306–317.

Yoneyama, N., Crabbe, J. C., Ford, M. M., Murillo, A., & Finn, D. A. (2008). Voluntary ethanol consumption in 22 inbred mouse strains. *Alcohol*, *42*(3), 149–160. https://doi.org/10.1016/j.alcohol.2007.12.006

Appendix 1: Files and Data

These files are presented as either dropbox links or pathways to the file on the VIPBG group server. The Dropbox links will take you to the folder where all of the listed files can be found.

Dropbox Links

Data Preparation

Count Data Preparation

B6 Aligned Counts

D2 Aligned Counts

Differential Expression Analysis

Working Directory

Results

B6 Aligned

D2 Aligned

Comparisons

Differential Exon Usage Analysis

B6 Aligned DexSeq

D2 Aligned DexSeq

Comparison of Results

Paths to Group Server Files

/home/projects/MilesLab/teamshare/DZ_B6_Alignment/

/home/projects/MilesLab/teamshare/DZ_D2_Alignment/

/home/projects/MilesLab/teamshare/DZ_D2_Alignment_DexSeq/

/home/projects/MilesLab/teamshare/B6D2_DeepSeq/

Appendix 2: Code

This appendix follows the same format as the previous. Dropbox links to folders that contain the scripts used, and paths to the scripts on the group server.

Dropbox Links

Count Data Preparation

"C:\Users\zelif\Dropbox (MilesLab)\Miles and Dustin Z\Aim 1 - Differential Exon, Differential Expression, Gene Ontology, and Transcript Level\Data Preparation\Count Data Preparation\commonkeys_first_step_of_gene_ID_conversion.py"

"C:\Users\zelif\Dropbox (MilesLab)\Miles and Dustin Z\Aim 1 - Differential Exon, Differential Expression, Gene Ontology, and Transcript Level\Data Preparation\Count Data Preparation\Gene ID and Name extraction from gff3 script.py"

Differential Expression Analysis

"C:\Users\zelif\Dropbox (MilesLab)\Miles and Dustin Z\Aim 1 - Differential Exon, Differential Expression, Gene Ontology, and Transcript Level\Differential Expression Analysis\Code\DZ_DESeq2_B6_aligned_script_3_19_23.R"

"C:\Users\zelif\Dropbox (MilesLab)\Miles and Dustin Z\Aim 1 - Differential Exon, Differential Expression, Gene Ontology, and Transcript Level\Differential Expression Analysis\Code\GTF Conversion for DEX seq.py"

Differential Exon Utilization Analysis

Paths to Group Server Scripts

/home/projects/MilesLab/teamshare/DZ_D2_Alignment/scripts/

/home/projects/MilesLab/teamshare/DZ_D2_Alignment_DexSeq/