Computational study of the effects of early life trauma on gene expression and exon usage in various tissues and cells in *Mus musculus*

A Thesis
Presented to
The Division of Biosystems Science and Engineering
of the ETH Zürich, July 2019

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Computational Biology and Bioinformatics

---

Andrew Stephen Acciardo

Supervision by: Deepak K. Tanwar and Prof. Dr. Isabelle Mansuy

Approved for the Division
(Biosystems Science and Engineering)

_____

Prof. Dr. Isabelle Mansuy

Acknowledgements

I would first and foremost like to thank Prachi for giving me so much support and love over the past 3 years. Through both the good and bad times you were always there to cheer me on and help me whenever I needed it. There were many points were I was not sure if we would make it out of this program, but we did it! Juno, my wonderful cat, deserves my thanks as well. In the most stressful and dire moments of my Master's studies, Juno was there to make me laugh and allowed me to enjoy the simple things in life. I also want to thank Deepak for mentoring me and guiding me through my project since January. It wasn't always easy and I appreciate your patience and willingness to answer even my most simplistic questions! You taught me a lot and I will take everything I learned from you to my future endeavors. Additionally, I want to thank Pierre-Luc for all of his critical help and assistance throughout my project. When all else failed, I could rely on you to find the best solution to any given problem, or to look at it in a new way. Finally I would like to thank Professor Mansuy for allowing me to work in such a wonderful laboratory as well as to get the chance to work on a project that challenged me and allowed me to learn so many new skills.

Table of Contents

Abstract

RNA-Seq technologies and analysis techniques developed to handle this type of data have revolutionized the fields of molecular biology and genomics. In recent years, these breakthroughs have been leveraged to investigate the mechanisms of epigenetic inheritance and transgenerational epigenetic inheritance (TEI) in mammals. To study the mechanisms of epigenetic inheritance, data was obtained from seven RNA-Seq experiments and was subjected to extensive computational analysis. For this thesis, the aims were to identify the patterns of differential gene expression and exon usage in somatic and germline cells from the MSUS mouse (*Mus musculus*) model.

Differentially expressed genes from the seven datasets were associated with functional and cellular processes that are known to be affected in MSUS mice. Extensive differential gene expression was seen in the spermatogonial cell (SC) adult and sperm datasets and widespread differential exon usage was found in the SC PND8 and Tesa1-day datasets. Furthermore, over 100 KEGG pathways were found to be significantly enriched in both the liver F3 and sperm datasets. The altered pathways have involvement in metabolic, neurological, and transcriptional regulation processes. These analyses provide a solid foundation for multi-dataset and multi-omics integrative analysis and for experimental validation of MSUS-related genes and biological pathways.

List of Abbreviations

- BH - Benjamini-Hochberg
- CPM - Counts-per-million
- DEG - Differentially expressed gene
- DEU - Differential exon usage
- DGE - Differential gene expression
- F0 - Filial generation 0
- F1 - Filial generation 1
- F3 - Filial generation 3
- F4 - Filial generation 4
- F5 - Filial generation 5
- FCS - Functional class scoring
- FDR - False discovery rate
- GLM - Generalized linear model
- GSEA - Gene set enrichment analysis
- HPA - Hypothalamic-pituitary-adrenal axis
- IDE - Integrated Development Environment
- MSUS - Unpredictable maternal separation combined with unpredictable maternal stress
- NB - Negative binomial
- ORA - Over-representation analysis
- PCA - Principal Component Analysis
- PND - Postnatal day
- PT - Pathway topology
- PTM - Post-translational modification
- SC - Spermatogonial cell
- TEI - Transgenerational epigenetic inheritance
- TMM - Trimmed mean of M values

Chapter 1

Introduction

1.1 Epigenetics

Changes in the gene expression patterns in organisms that cannot be attributed to genomic alterations are classified under the term "epigenetics" (Waddington, 1942). Taken as a whole, these changes make up what is known as the epigenome. In contrast to the genome, the epigenome is dynamic, variable, and can be markedly different between cells and tissues in the same organism, depending on a given time point or condition (Suzuki & Bird, 2008). Epigenetic modifications are critical for both the differentiation and the development of complex multicellular life forms. Also, they are necessary for the standard moment-to-moment maintenance and processes that are involved in keeping the organism functioning (Skvortsova, Iovino, & Bogdanović, 2018). The disruption of the functioning of epigenetic processes has been demonstrated to be linked to various diseases and disorders including Alzheimer's disease, schizophrenia, diabetes, autoimmune disorders, and multiple forms of cancer (Allis & Jenuwein, 2016; J. D. Sweatt, 2013).

1.2 Neuroepigenetics

The study of epigenetic changes in the nervous system has increased substantially over the past 15 years (J. D. Sweatt, 2013). During this time, epigenetic mechanisms have been suggested and observed to have some impact on a variety of human nervous system functions including stress responses (Matosin, Cruceanu, & Binder, 2017), learning and memory (Zovkic, Guzman-Karlsson, & Sweatt, 2013), and Alzheimer's disease (Nativio et al., 2018). The key process of regulation of the central stress response occurs along what is known as the hypothalamic-pituitary-adrenal (HPA) axis (Dick & Provencal, 2018). Exposure to psychological trauma has been shown to introduce stable epigenetic changes in genes that are part of the HPA axis, leading to either hypo-expression or hyper-expression of key genes which consequently results in altered capabilities for handling stress (Provençal & Binder, 2015).

1.3 Epigenetic modifications

The primary forms of epigenetic modifications include covalent modifications of DNA, the most important of which is methylation of cytosines at CpGs, as well as post-translational modifications (PTM) of histones, and the actions of non-coding RNAs including microRNAs (miRNA) (A. Bird, 2002; J. D. Sweatt, 2013). Cytosine methylation, in particular, plays a critical role in regulating the accessibility of DNA. When CpG islands, located in the promoter regions of DNA, are methylated, transcription factor binding is inhibited and consequently, gene expression is repressed. The distribution of methylated cytosines throughout the mammalian genome is an important factor in maintaining the stability of the genome. Studies have shown that the dysregulation of cytosine methylation can lead to cell and organismal death (Li, Bestor, & Jaenisch, 1992; Panning & Jaenisch, 1996).

Histone modifications represent another key method of epigenetic control. Within the nucleus of all eukaryotic cells, DNA wraps around octamers of proteins called histones, forming nucleosomes. Histones help condense DNA and play a role in gene regulation. Importantly, the structure of histones includes a tail which protrudes out of the nucleosome and is a target for phosphorylation, acetylation, and methylation, among many other possible modifications. These modifications can either facilitate or hinder access to the DNA. Similarly to DNA methylation, histone modifications have been implicated in human diseases and disorders including Coffin-Lowry syndrome (Hanauer, 2002) and Kleefstra syndrome (Benevento et al., 2016).

Recently, epigenetic alterations originating from the actions of non-coding RNAs (ncRNA) have been studied with increasing interest. The most active area of research is microRNAs (miRNA), which are known to modulate gene expression both by targeted degradation of mRNA and by binding to DNA, thereby blocking transcription (Huntzinger & Izaurralde (2011)]. Within the brain, miRNA appears to be abundant and have been shown to have a clear role in the brain's ability to respond to acute stress (Mannironi et al., 2013).

1.4 Epigenetic Inheritance

Despite the reprogramming of the epigenome in germ cells and embryos in mammals, the capability of epigenetic alterations to be transmitted to the offspring across generations has been well documented (Lacal & Ventura, 2018; Lind & Spagopoulou, 2018; Skvortsova et al., 2018). Intergenerational inheritance is defined as the transmission of epigenetic marks from parent to direct offspring (F0 to F1), while transgenerational inheritance involves transmission from grandparents to grand-offspring and to further generations (F1 to F3, F4, and F5) (Skvortsova et al., 2018). The process of epigenetic inheritance can begin when an organism is influenced in some way by its environment, which then causes changes in the epigenome of gamete cells. When one of these affected gametes is used in the formation of a zygote, the epigenomic alterations can be transferred to the offspring, leading to altered transcriptional regulation and consequently altered phenotypes in the offspring. The manifestations of these epigenetic changes in the offspring can lead to further propagation to subsequent generations. While the exact mechanisms of how the germ cell perturbations affect the transgenerational inheritance of behavioral and metabolic alterations are still not known, it is an active area of research (Bohacek

& Mansuy, 2015).

1.5 Unpredictable maternal separation combined with unpredictable maternal stress (MSUS)

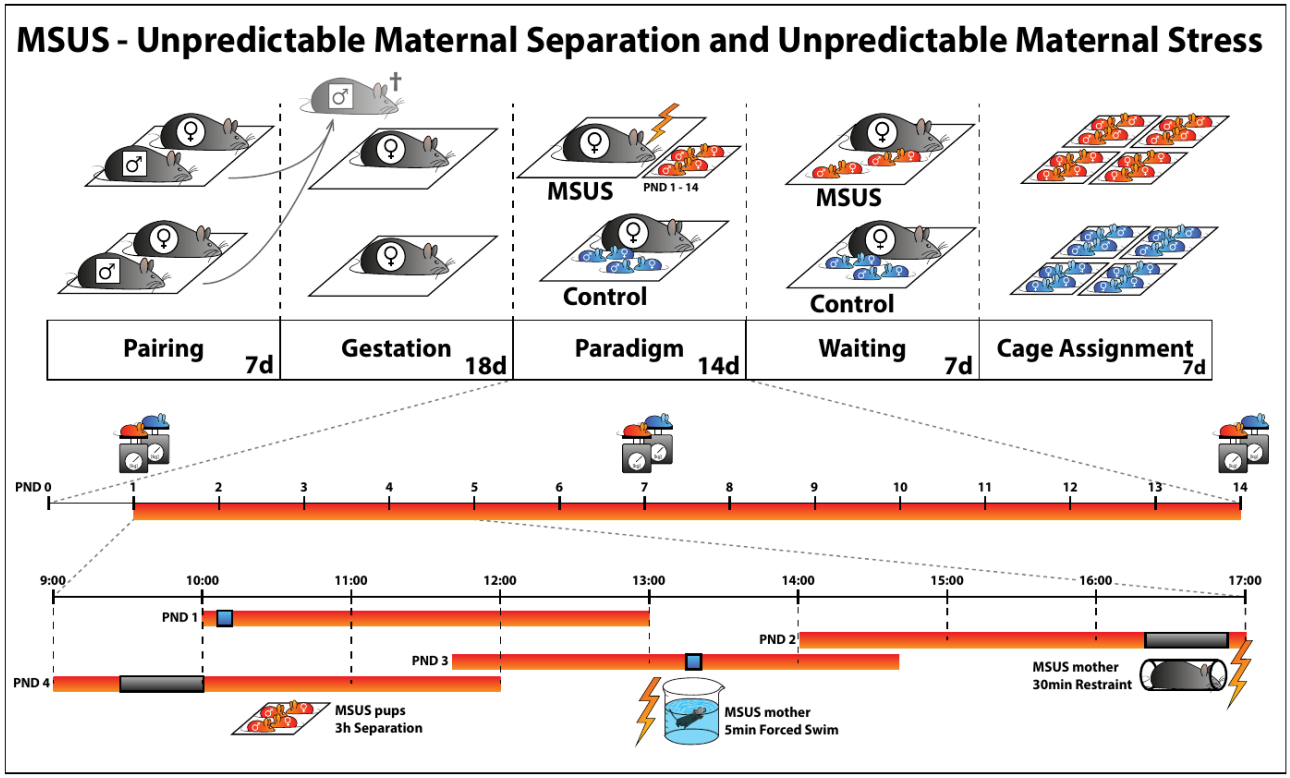

Figure 1.1: Unpredictable Maternal Separation Combined With Unpredictable Maternal Stress (MSUS) model - Image was taken from Martin Roszkowski, a PhD student in the Mansuy Laboratory.

Traumatic experiences have been shown to produce changes in organisms that are inheritable across generations (T. B. Franklin et al., 2010). However, not all aspects of these changes can be attributed to genetic etiologies. One suggestion was that epigenetic mechanisms were enabling the transmission of these alterations. To test this hypothesis, a mouse (*Mus musculus*) model was designed that involved subjecting newborn pups (between postnatal day (PND) 1 to 14) to chronic and unpredictable maternal separation, as well as subjecting the mother to unpredictable stressors. Hence, the model is called unpredictable maternal separation combined with unpredictable maternal stress (MSUS) (Figure 1.1).

Briefly, during PND1-14, pups are separated at unpredictable times from their mother for a period of 3 hours each day. Concurrently to the separation, the mother is subjected to certain stressors including a 5 minute forced-swim in the water, or restraint in an enclosed space for 30 minutes. The intuition behind the MSUS model is that by introducing the unpredictability into the maternal separation, the mother is unable to anticipate the separation and compensate for

it. Similarly, the unpredictability of the stressful situations, which the mother is subjected to, serves to increase the effect of the diminished maternal care provided to the pups. After the MSUS treatment is complete, both the MSUS pups and control pups are tested with a variety of methods, including navigating an elevated maze (Walf & Frye, 2007) and a forced swim test (Can et al., 2011). The tests are meant to determine whether any significant changes in behavior, metabolism or otherwise had become apparent in the MSUS pups. Pups subjected to the MSUS paradigm display depressive-like behaviors and alterations in DNA methylation profiles. It is also shown that these changes were present in the offspring of males subjected to this paradigm, even though the offspring themselves had not experienced this trauma. (T. B. Franklin et al., 2010). Subsequent studies expanded on this work by demonstrating that the MSUS paradigm can also cause changes in metabolism (Gapp et al., 2014a), alterations in oxytocin and cortisol (Babb, Carini, Spears, & Nephew, 2014), and impaired spatial memory (Bohacek & Mansuy, 2015). Importantly, certain metabolic changes have been seen even in the fourth generation after the initial trauma (Steenwyk, Roszkowski, Manuella, Franklin, & Mansuy, 2018).

Not all of the changes induced by the MSUS appear to be negative. Gapp et al. showed that adult offspring of male pups subjected to the MSUS paradigm demonstrated an affinity for goal-directed behavior and higher behavioral flexibility (Gapp et al., 2014b). These findings are of note because it offers evidence of the adaptability and resilience of these animals in the face of adversity, showing that they are better able to cope with difficult situations later in life, which could have significant implications for the survival and propagation of a species.

1.6 RNA-Seq

The advent of high-throughput sequencing technologies in the past two decades laid the foundation for an entirely new field of research and allowed researchers to explore questions about life and its many facets at a level of detail that was previously impossible (Z. Wang, Gerstein, & Snyder, 2009). In order to study the phenotypic and genotypic variation that is present in eukaryotic organisms, the technique of transcriptomics has been developed. This technique looks at the transcriptome, which is the complete set of transcripts present in a cell at a given point in time. An early method of interrogating the transcriptome was hybridization-based microarrays, which work by taking the mRNA from cells, tagging them with a fluorescent marker, converting them to cDNA, and placing them on the array to see if they bind. (Bumgarner, 2013).

As deep-sequencing technologies continued to develop in the mid-to-late 2000s, the ability to map and quantify transcriptomes quickly and accurately brought about the creation of RNA-Sequencing (RNA-Seq) methods. Today, the standard high-throughput sequencing platform in use is Illumina sequencing, which uses bridge amplification to create clonal clusters of each cDNA, which are then sequenced "by synthesis" (Berge et al., 2018). The advantages of RNA-Seq are numerous, among the most important of which is that it can be used to construct genomes and transcriptomes de novo, without the need of any reference sequence. In contrast to microarrays, which can be hindered by background signal due to cross-hybridization or other issues, RNA-Seq does not have this problem, because it is able to map sequences unambiguously and with base-pair resolution (Z. Wang et al., 2009). Figure 1.2 shows the basic steps of RNA-Seq as well as subsequent analyses that can be performed on RNA-Seq data.

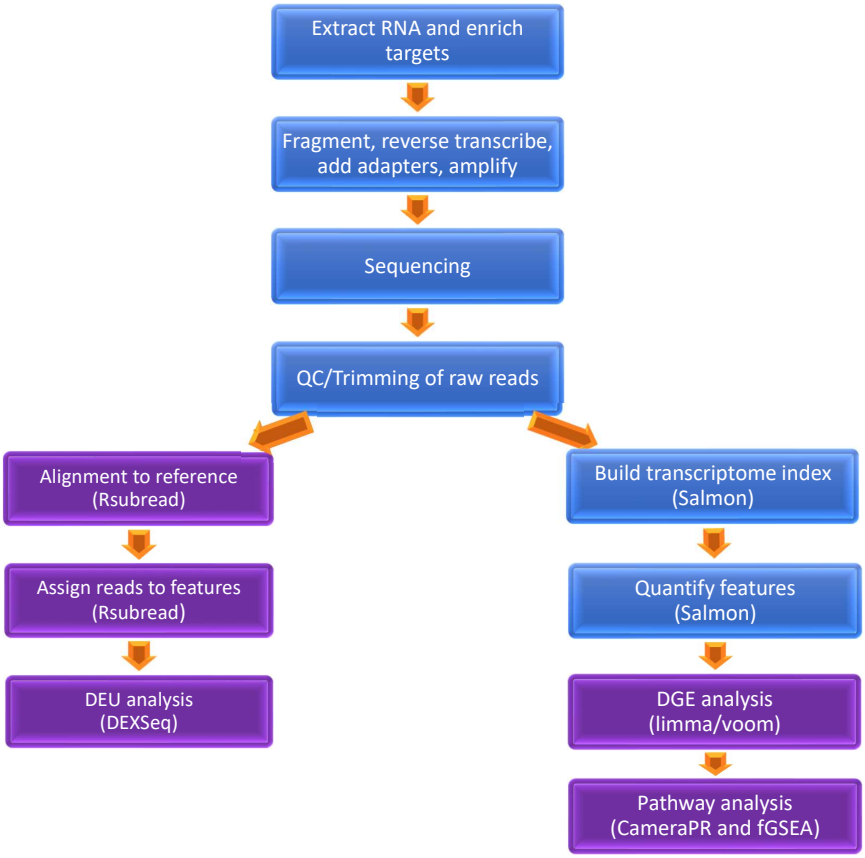

Figure 1.2: RNA-Seq and downstream analysis workflow: Flowchart detailing the steps involved in RNA-Seq and subsequent analyses. Steps in purple were completed in this thesis work.

1.7 Alignment and Quantification

Raw reads from deep-sequencing runs must go through quality control and trimming steps to check for GC or nucleotide composition bias, sequence length distributions, sequence duplications, and to remove poor quality reads. After this, the reads are mapped to a reference genome and the resulting counts of reads that overlap with genes/transcripts are quantified. The plethora of tools used to perform read counting can be roughly split into two major groups; namely, those that must do full read alignments before quantification, and those that are capable of alignment-free quantification (Berge et al., 2018). Rsubread, an R package, belongs to the first group. This package contains methods for both mapping and quantification, using a seed-and-vote method for mapping and an R implementation of featureCounts for read summarization (Liao, Smyth, & Shi, 2019). Before Rsubread can be used, a hash table index of the reference genome must be created. In contrast to the direct alignment method of Rsubread, Salmon has the ability

to create and use an index to quasi-map RNA-Seq reads, meaning that they are assigned to reference locations rather than doing base-to-base alignment (Patro, Duggal, Love, Irizarry, & Kingsford, 2017). This strategy of alignment-free quantification is primarily advantageous due to greatly reduced computational requirements and faster running time.

1.8 Differential Expression Analysis

In general, the end result of alignment and quantification is a matrix of counts for the genes/transcripts of each sample in the experiment. From this point, it is possible to perform differential gene expression (DGE) analysis, which aims to detect statistically significant changes in expression levels of genes between experimental conditions. There are a few approaches to DGE analysis, with the most widely used packages being edgeR, DESeq2, and limma/voom. Steps in DGE analysis include filtering and normalization of gene counts, specifying a model to be fitted to genes, estimation of parameters, inference, and multiple testing corrections. For the statistical modelling and estimation steps for RNA-Seq data, limma uses a linear model-based approach and assumes that the underlying data is normally distributed. Moderation of the dispersion estimates is done by sharing information across genes, using an empirical Bayes approach.

In order to check for differential expression, limma fits linear models to genes, which allows for the entire experiment to be analyzed together using linear regression (Ritchie et al., 2015). However, these linear models assume equal error variance, which is sufficient for microarray analysis, but not for RNA-Seq as the error rises with the mean. The solution to this issue was offered by the authors of limma in the form of voom - a method of generating a precision weight value and using this in the limma pipeline (C. W. Law, Chen, Shi, & Smyth, 2014). Using the voom approach, the RNA-Seq read counts are first converted to $\log_2$-counts-per-million ($\log_2$CPM) and then voom is used to model the mean-variance relationship with the precision weights. At this point, the data is able to go through the rest of the limma pipeline.

1.9 Differential Exon Usage Analysis

One gene can potentially encode many different isoforms. This process, known as alternative splicing, greatly increases the functional diversity of genes and presents many challenges in the context of RNA-Seq analysis. Over time, various approaches have been developed to analyze this phenomenon, including differential transcript expression which measures the difference in expression of individual transcripts between groups (Trapnell et al., 2012). Differential transcript usage, on the other hand, looks at equally expressed genes that show variable isoform abundance (Froussios, Mourão, Simpson, Barton, & Schurch, 2017). Finally, there is differential exon usage (S. Anders, Reyes, & Huber, 2012). Differential exon usage analysis tries to answer the question of whether or not the relative usage of a particular exon within a gene is different based on experimental conditions.

Currently, there are not many packages or software capable of performing this specific type of alternative splicing analysis. However, the DEXSeq R package is an exception (S. Anders

et al., 2012). DEXSeq calculates the ratio of the number of transcripts from a particular gene that contain an exon of interest to the total number of transcripts from the gene. DEXSeq implements generalized linear models (GLM) to model read counts, which also allows for the handling of complex experimental designs. It assumes a negative binomial distribution, in order to account for biological variation. This is in contrast to earlier methods in differential analysis software that used the Poisson distribution which is insufficient for RNA-Seq data (M. D. Robinson, McCarthy, & Smyth, 2009). By analyzing the read densities throughout each gene, DEXSeq produces a P value for differential usage for each exon, along with its associated fold change between conditions.

1.10 Pathway Analysis

The goal of pathway analysis is to analyze genes against a set of pathways with known interactions, and seeing which pathways are most up-regulated or down-regulated (Khatri, Sirota, & Butte, 2012). This is in contrast to gene set analysis, which uses gene groups that may not necessarily be structured based on known interactions. Using differential gene expression results, pathway analysis is used to determine whether there are any biologically relevant processes affected. According to Khatri et. al., there are three generations of pathway analysis methods.

The first of these is the over-representation analysis (ORA). As it is the earliest type of method developed, it is also the most simplistic. Given a list of gene sets, a threshold value of differential gene expression, and an FDR, ORA tests the fraction of genes in a given pathway, among all genes that passed the chosen differential expression and FDR thresholds. In ORA, every pathway is tested for over-representation or under-representation. A significant limitation of ORA methods is that it only considers the number of genes, but not the value (e.g. fold change) associated with them. This characteristic ignores critical information that could improve the accuracy of the analysis. Additionally, ORA only considers genes that are deemed differentially expressed based on the arbitrary threshold values that the user provides. Genes that just barely miss this threshold might still be significant, and so leaving them out may reduce the quality of the results due to the loss of information. Another weakness of ORA is that it assumes both gene and pathway independence, which is almost never the case in reality. Approaches that use ORA include GenMAPP 2 (Salomonis et al., 2007) and Genemerge (Castillo-Davis & Hartl, 2003).

The second is functional class scoring (FCS). The idea behind FCS is that a gene-level statistic (i.e q-statistic, z-score, and t-test) is calculated for each gene in a pathway and then these statistics are aggregated into a pathway-level statistic, which is then tested for significance. One improvement of FCS over ORA is the ability for FCS methods to account for interdependencies among genes. Among FCS methods, one can perform either competitive tests or non-competitive tests. Competitive tests compare genes in test set relative to all other genes, while non-competitive tests compare genes within a pathway with one another. Similarly to ORA methods, FCS methods still assume pathway-pathway independence and also, the fold change value associated with each gene is not considered after ranking genes. Two commonly used FCS methods include Camera (Wu & Smyth, 2012), from the limma R package, and GSEA (Subramanian et al., 2005), which was adapted to create the fGSEA package (Sergushichev, 2016). Camera and fGSEA test the ranking of a gene set relative to the other genes in the

experiment, taking correlation into account. fGSEA works similarly to the popular GSEA method, except that it calculates a cumulative gene set enrichment statistic value from a single sample, whereas GSEA performs sampling of a large number of gene sets simultaneously.

The third and most recent generation of pathway analysis techniques is pathway topology-based methods (PT). The steps of PT are for the most part identical to FCS, however PT is able to integrate topological data (i.e how genes interact with one another and where genes interact with one another) from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome databases to assist in calculating the gene-level statistic values, which can lead to greater sensitivity and specificity (Khatri et al., 2012). In general, PT is not able to be utilized in many cases, as these databases are still not rich enough to include pathway topologies for all cell-specific and condition-specific situations. Current PT approaches include GGEA (Geistlinger, Csaba, Kuffner, Mulder, & Zimmer, 2011) and SPIA (Tarca et al., 2008).

1.11 Aim

The overall aim of this project was to elucidate the transcriptional mechanisms underlying epigenetic inheritance in mice and to identify key genes and pathways that are altered in the MSUS paradigm using RNA-Seq data. The specific aims of the project were:

1. To characterize the differential gene expression found in mice either exposed to the MSUS paradigm or injected with a PPAR agonist that mimics certain MSUS effects.
2. To use the differential gene expression data to determine which, if any, biological pathways were affected
3. To determine the extent to which the differential exon usage was present in MSUS or PPAR agonist-injected mice

Chapter 2

Methods

2.1 Datasets

Seven RNA-Seq datasets were analyzed for this study. All samples from each dataset derived from male mice. A summary of each dataset can be found in Table 1.

Table 1: Dataset information

Dataset	n	Experimental n	Control n	Generation	Tissue	Type
Somatic + Testis	21	11	10	F1,F3	Hippocampus,Liver,Testis	Somatic,Testis
SC Adult	12	6	6	F1	Spermatogonia	Germline
SC Pups	32	15	17	F1	Spermatogonia	Germline
Sperm	13	7	6	F1	Sperm	Germline
Zygote	8	4	4	F1	Zygote	Germline
Tesa46-day	14	7	7	F1	Sperm	Tesaglitazar
Tesa1-day	8	4	4	F1	Sperm	Tesaglitazar

2.1.1 Hippocampus, Liver, and Testis

These data were generated by Martin Roszkowski, a PhD student in the Mansuy laboratory. In total, there were 63 samples from 21 mice from the MSUS23 breeding. RNA-Seq was performed on three different tissues from each mouse - the liver, in order to check for metabolic changes which have been seen in MSUS paradigms; the hippocampus, to check for possible CNS changes since the MSUS paradigm leads to behavioral alterations; and the testes. The samples from this dataset include mice from the F1 generation, as well as their direct descendants from the F3 generation, which was done to enable comparisons between the two generations and to see if alterations to the F1 mice were transmitted to the F3 mice. The set of 21 mice was composed of 10 CTRL samples and 11 MSUS samples. Samples were sequenced at the Beijing Genomics Institute (BGI) (Beijing, China) using a proprietary BGI sequencing platform. The project is available at `https://github.com/mansuylab/MSUS23_tissue_RNASeq`

2.1.2 Spermatogonial Cells (SC) - Adult

These data were generated by Irina Lazar-Contes, a PhD student in the Mansuy laboratory. In total, there were 12 samples from the MSUS30 breeding. RNA-Seq was performed on spermatogonial cells sorted from adult F1 mice (approximately 5 months old). The set of 12 mice was composed of 6 CTRL samples and 6 MSUS samples. The library preparation kit used was Nextera XT. Sequencing was done on the Illumina NovaSeq at the Functional Genomics Center Zurich. The project is available at `https://github.com/mansuylab/SC_adult`

2.1.3 Spermatogonial Cells - PND8/PND15

These data were generated by Irina Lazar-Contes. In total, there were 37 samples from 32 mice from the MSUS27 breeding. RNA-Seq was performed on spermatogonial cells sorted from F1 pups at PND8 (during MSUS treatment) and PND15 (after MSUS treatment). The set of mice was composed of 17 CTRL samples and 15 MSUS samples. Sequencing was done on the Illumina NovaSeq at the Functional Genomics Center Zurich. Certain control samples were re-sequenced due to low number of reads - these samples are denoted with "mRNA" in their sample name. The project is available at `https://github.com/mansuylab/SC_longRNA`

2.1.4 Sperm

These data were generated by Dr. Katharina Gapp, a former member of the Mansuy laboratory. In total, 13 samples were sequenced. RNA-Seq was performed on sperm cells sorted from F1 mice. The set of 13 mice was composed of 6 CTRL samples and 7 MSUS samples. The library preparation kit used for some samples was Nextflex, while for others it was Illumina TruSeq. Sequencing was done on the Illumina HiSeq 2500 at the Sanger Institute (Cambridge, United Kingdom). The project is available at `https://github.com/mansuylab/sperm_kathi`

2.1.5 Tesa46-day

These data were generated by Dr. Gretchen van Steenwyk, a post-doctoral researcher in the Mansuy laboratory. In total, 14 samples were sequenced. RNA-Seq was performed on sperm cells sorted from F1 mice. Rather than going through the canonical MSUS paradigm, mice were injected either with saline (CTRL) or tesaglitazar (TESA), a PPAR agonist, in order to try to artificially replicate the metabolic effects of MSUS. Mice were injected with either saline or tesaglitazar 2 times per week for 4 weeks. Sperm samples were collected 46 days after the injections ended. The set of 14 mice was composed of 7 CTRL samples and 7 TESA samples. The library preparation kit used was Illumina TruSeq. The project is available at `https://github.com/mansuylab/tesa_sperm_1`

2.1.6 Tesa1-day

These data were generated by Dr. Gretchen van Steenwyk. In total, 8 samples were sequenced. RNA-Seq was performed on sperm cells sorted from F1 mice. As in the Tesa46-day dataset, the mice were injected either with saline or tesaglitazar 2 times per week for 4 weeks. Sperm samples were collected 1 day after the injections ended. The set of 8 mice was composed of 4 CTRL samples and 4 TESA samples. The library preparation kit used was Illumina TruSeq. Samples were sequenced with Illumina NovaSeq. The project is available at `https://github.com/mansuylab/tesa_sperm_2`

2.1.7 Zygote

These data were generated by Dr. Katharina Gapp. In total, 8 samples were sequenced. RNA-Seq was performed on zygotes sorted from F1 mice. The set of 8 mice was composed of 4 CTRL samples and 4 MSUS samples. The library preparation kit used was Illumina TruSeq. The project is available at `https://github.com/mansuylab/zygote_kathi`

2.2 Code development

All work was done in a UNIX environment (Ubuntu v16.1). Statistical analyses and manipulation of data sets and data structures were done using the R programming language, version 3.5. The integrated development environment (IDE) RStudio, version 1.1, was used to write and develop scripts that would be used to run analyses.

2.3 Rsubread

Alignment and quantification for DEXSeq were done using the Rsubread package. First, an index of the *Mus musculus* genome (GRCm38.p5) was constructed using the `buildindex()` function. Then the mapping of trimmed reads was performed with the `subjunc()` function, which is recommended over the `align()` function, due to its additional functionality of detecting exon-exon junctions. The output format was set to BAM, the phredOffset was set to 33, and the Gencode vM18 chr_patch_halp_scaff annotation was used. The final step was to provide the `featureCounts()` function with the output from the alignment step, to perform quantification. For this step, the allowMultiOverlap option was set to TRUE, in order to allow a read to be assigned to multiple features if it overlaps with more than one. Additionally, primaryOnly was set to TRUE, so that only primary alignment were counted. Finally, the strandSpecific option was set to 2, which made featureCounts perform reverse-strand specific counting since the libraries were prepared to be reverse-stranded.

2.4 DEXSeq

To perform differential exon usage (DEU) analysis, the R package DEXSeq was used. The resulting count data from Rsubread alignment was used to test for DEU. The first step was loading the data and creating a DEXSeqDataSet object using count files that contain the number of mapped reads for each exon, a file containing meta-data about the samples, a design formula that determines which differences in exon usage will be examined, and a flattened GFF annotation file. Next, the samples were normalized using the median ratio method (Simon Anders & Huber, 2010) in order to adjust for different sequencing depths. Then, to estimate the variability of the data, dispersion estimates were calculated by using a Cox-Reid adjusted profile likelihood shrinkage approach (Love, Huber, & Anders, 2014). To check for differential exon usage, DEXSeq implements a likelihood ratio test to test a full model with an *exon x condition* interaction term against a reduced model without the *exon x condition* interaction term. This allows for one to see if there is differential exon usage that is solely attributable to the MSUS paradigm. Exons with an FDR value of less than 0.1 were selected. A generalized R script was developed that is capable of taking input from any of the seven data sets studied in this work and performing differential exon usage analysis, with various options for the comparison to be made, effects to correct for, tissues or samples to subset, and the FDR. The output of the script is an .html file generated by the `DEXSeqHTML()` function, which provides a browseable report containing all relevant results and plots.

2.5 edgeR/limma/voom/fdrtool

Differential gene expression analysis was performed using the edgeR and limma R packages. Gene counts from Salmon were first filtered by removing any gene where more than 40% of the samples had a count value of less than 15. This was done because RNA-Seq measurement errors are worse in lowly expressed genes and this can negatively impact the power to detect differentially expressed genes. After filtering, normalization was carried out on the counts using the `calcNormFactors()` function from edgeR, which performs trimmed mean of M values (TMM) normalization. Since limma was originally designed to work with microarray data, "voom" normalization was carried out on the normalized and filtered gene counts to calculate a precision weight that was used to model the mean-variance relationship. The vector of t-statistics obtained from the limma/voom analysis was then inputted into the `fdrtool()` function, from the fdrtool R package, in order to re-estimate the null model and recalculate P values. This was done for cases when the original P value histograms had distributions that were not uniform. For example, when the histogram distribution was strongly skewed towards high P values, this typically meant that the variance of the null-distribution was overestimated. The fdrtool algorithm first takes the test statistics and fits an approximate null model, after which a cutoff point with the smallest possible false non-discovery rate is determined. Next, the estimates for the proportion of null values and scale parameters are obtained. From these estimates, "corrected" P values are computed and saved.

2.6 Camera pre-ranked

Pathway analysis was performed with two separate R packages/methods. The first of these is CameraPR, which is a function that is part of the limma package. "Pre-ranked" means that the t-statistic from the limma differential gene expression analysis was taken for each gene, and the genes were ranked based on the value of the t-statistic. All genes, including those that were not selected as significant, were ranked and included in the CameraPR analysis. Pathways with an FDR value of less than 0.1 were selected. The original `camerapr()` function was modified to report the t-statistic of the CameraPR analysis, in order to be used in subsequent visualization steps.

2.7 fGSEA

The second pathway analysis method used was fast Gene Set Enrichment Analysis (fGSEA). This was done with the fGSEA R package. Similarly to Camera, this method used pre-ranked genes to run the analysis. The t-statistic value for each gene obtained during differential expression was used to calculate the ranks. The minimum size of a gene set to test was set at 10 in order to disregard small gene sets, which are more difficult draw biological conclusions from. The maximum gene set size was set at 5000 to ensure that all other gene sets would be included. The number of permutations was set at 1000, meaning that 1000 independent samples were generated for each gene set. A number higher than 1000 would likely not result in many more significant pathway hits.

2.8 Gene sets

Gene sets from *Mus musculus* for pathway analysis were obtained from two sources, the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome. The latest KEGG pathways were obtained using the `getGenesets()` function from the EnrichmentBrowser R package. The `getGenesets()` function returns the results in the form of Entrez Gene IDs, so in order to give the genes a more readable name, they were converted to their respective gene symbols using the `lookUp()` function from the annotate R package. In total, 329 gene sets were retrieved from KEGG. Reactome pathways were obtained by downloading from the R script directly from the official Reactome website. In total, 2231 gene sets were retrieved from Reactome. Results from Reactome pathway analysis are not described in this work.

Chapter 3

Results

3.1 Exploratory Data Analysis

For the purposes of comparative analysis, the samples and comparisons were separated into three groups. The first group contains somatic and testis samples from the Hippocampus, Liver, and Testis dataset. The testis samples may include both somatic and germline cells, so although they were meant to be part of the somatic group, they will be not be assumed to be so in this work. The second group is made up of germline samples from four datasets 1) adult spermatogonial cells 2) pup spermatogonial cells 3) sperm cells and 4) zygotes. The final group is made up of the two tesaglitazar sperm datasets. While the tesaglitazar injections were hypothesized to mimic some of the effects of MSUS treatment, tesaglitazar sperm samples were separated from the MSUS germline group, as the tesaglitazar experiments did not involve any aspects of the MSUS paradigm.

3.1.1 Normalization Plots

Distributions of gene expression counts before and after normalization were plotted to verify the efficacy of the procedure (Figure 3.1). Salmon counts were filtered to remove lowly expressed genes, after which factors to scale raw library sizes were obtained by TMM normalization, a method that accounts for sequencing depth and RNA composition by estimating relative RNA production levels between samples. Counts-per-million (CPM) normalization was then performed on the counts, taking into account the scaled library sizes, in order to facilitate gene count comparisons between all of the samples.

For somatic and testis samples, count distributions for all samples match up very well. In this case, all samples came from the same RNA-Seq run and the non-normalized count distributions already were nearly uniform. Germline sample count distributions for all samples also match up well. Samples originated from 4 different datasets, so the non-normalized count distributions varied more than the somatic and testis samples. Finally, tesaglitazar sperm sample count distributions for all samples match up. There were two separate datasets and the non-normalized count distributions were not uniform.

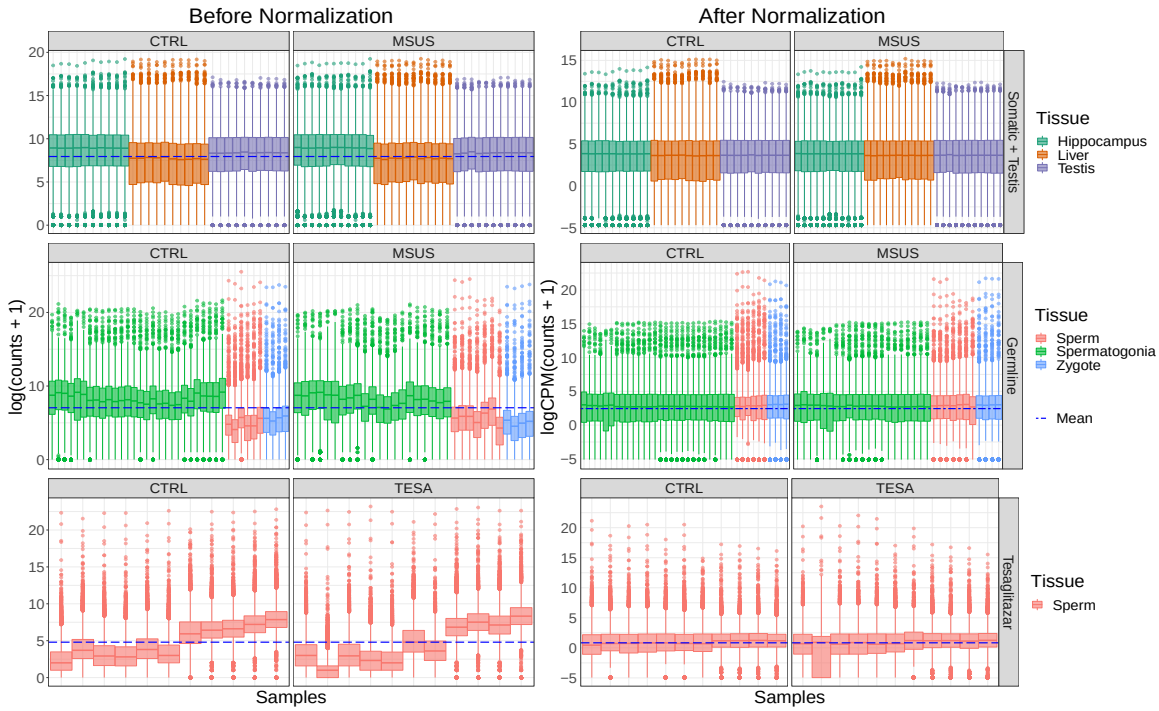

Figure 3.1: Normalization Plots - Box-plots of gene counts from all samples. Gene counts were obtained from the Salmon RNA-Seq quantification software. Counts were used to generate a DGEList object, after which lowly expressed genes were filtered and normalization factors to scale raw library sizes were calculated using the Trimmed Mean of M-values (TMM) method. Left: The $\log_2$ of the gene counts (Before Normalization) are shown for each sample. Right: The $\log_2$ counts-per-million (CPM) normalized counts (After Normalization) are shown for each sample. A blue dashed line in each plot denotes the overall mean value. The lower and upper hinges correspond to the first and third quartiles. The upper and lower whiskers extend from the hinge to the largest or smallest value no further than 1.5 * Inter-quartile range (IQR) from the hinge.

3.1.2 Principal Component Analysis

The principal component analysis was performed on the filtered Salmon gene counts to group together similar variables in the datasets. This helped to identify groups with similar gene expression profiles. The bottom panel of Figure 3.2 shows that the somatic and testis data very clearly separated into three groups based on the tissue they derived from. Within each tissue group, there does not seem to be any sub-grouping based on either generation or experimental condition. To determine the clustering of purely somatic samples, testis samples were from analysis (Figure 3.2. Unsurprisingly, the two clusters formed separate liver and hippocampus samples. Within both tissue groups, there does not seem to be any sub-grouping based on either generation or experimental condition. To check if the testis samples grouped with germline samples, they were included in the PCA of germline samples and plotted in Figure 3.3. While not quite as distant from the other samples as in the somatic PCA, the testis samples still form a separate group that is clearly distinct from the germline samples. The sperm and zygote datasets seemed to cluster closely together, as do the two spermatogonial cell datasets. Removing the testis samples and rerunning the PCA on the germline samples produced different results (top panel of Figure 3.3). Whereas in the previous PCA, the two spermatogonial cell datasets grouped together, they are clearly separated in the PCA without the testis samples. However, the sperm and zygote samples still appear to group together. For the two tesaglitazar sperm datasets, PCA results show two separate groupings based on the two datasets, although the Tesa46-day dataset seems to have moderately high within-dataset variability compared to the tesaglitazar sperm 2 dataset (Figure 3.4).

In all germline and tesaglitazar PCA plots, there does not seem to be any grouping of experimental conditions within any of the datasets.

3.2 Differential Gene Expression

The R package limma, along with voom, was used to perform differential expression analysis on the seven RNA-Seq datasets.

3.2.1 fdrtool *P* value Histograms

In order to visualize and inspect the distribution of DGE P values, histograms were plotted for each limma analysis of MSUS versus control samples. Additionally, the original P values were subjected to fdrtool correction to account for non-uniform distributions. In some cases, the original P value distribution did not necessarily require fdrtool correction. However, in order to keep the analysis consistent, the corrected P values were used from all comparisons for subsequent analysis. All somatic and testis comparisons, except for F1 testis, showed non-uniformities in their original P value distributions (Figure 3.5). The correction by fdrtool appears to have created more uniform distribution and significant P values are visible as a peak near zero on the histograms. The fdrtool corrected P values for SC Adult and SC PND15 have a

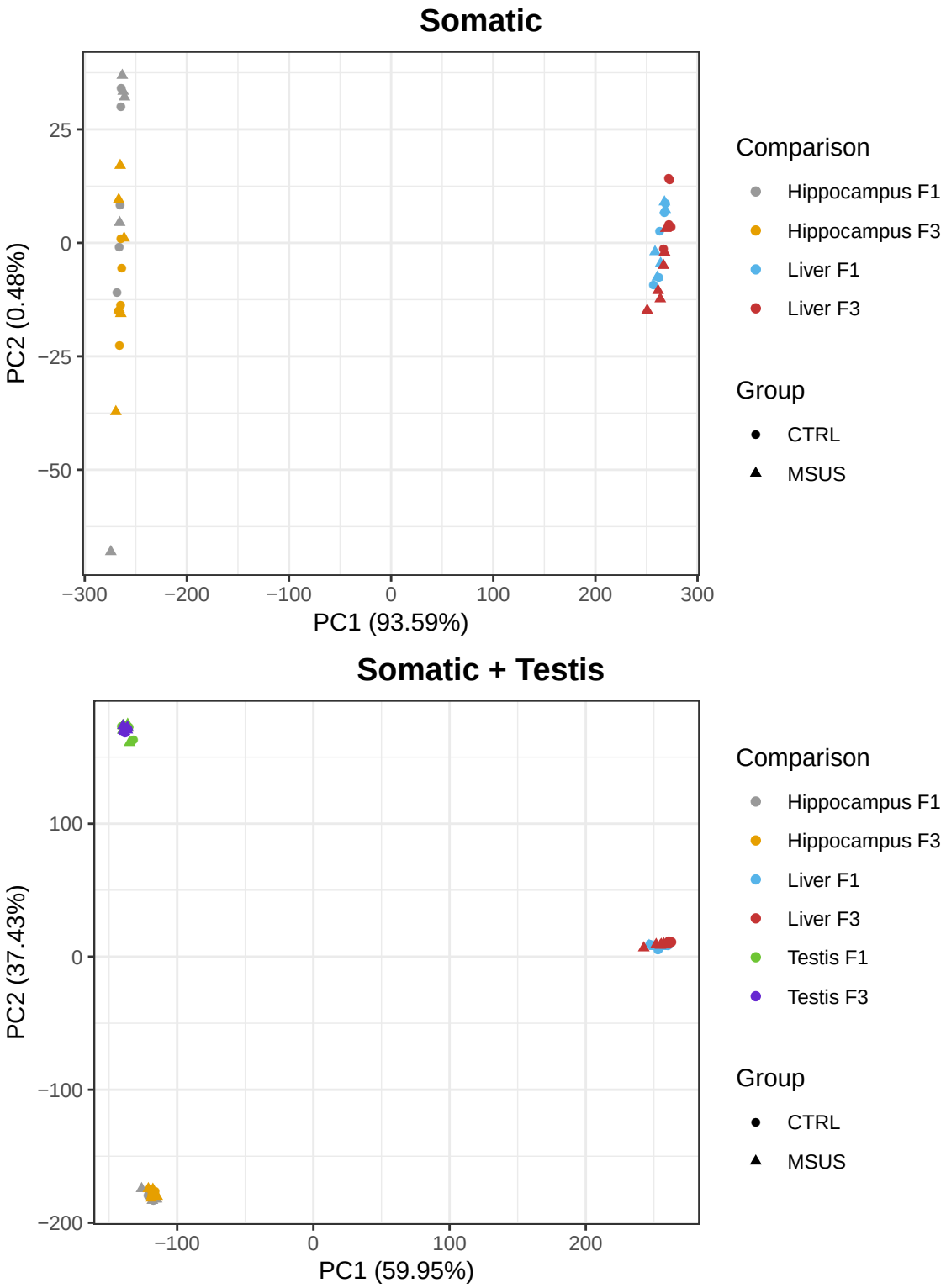

Figure 3.2: Principal Component Analysis - Somatic. PCA of gene counts from somatic and testis samples. Gene counts were obtained from the Salmon RNA-Seq quantification software. Lowly expressed genes were filtered and the remaining counts were normalized by $\log_2$CPM. Control samples are plotted as circles and MSUS samples as triangles. Different colors are used to differentiate both tissue and generation. Top: Samples coming from somatic tissues are plotted. Bottom: Samples coming from somatic and testis tissues are plotted. Samples grouped together based on tissue type, but not generation or group.

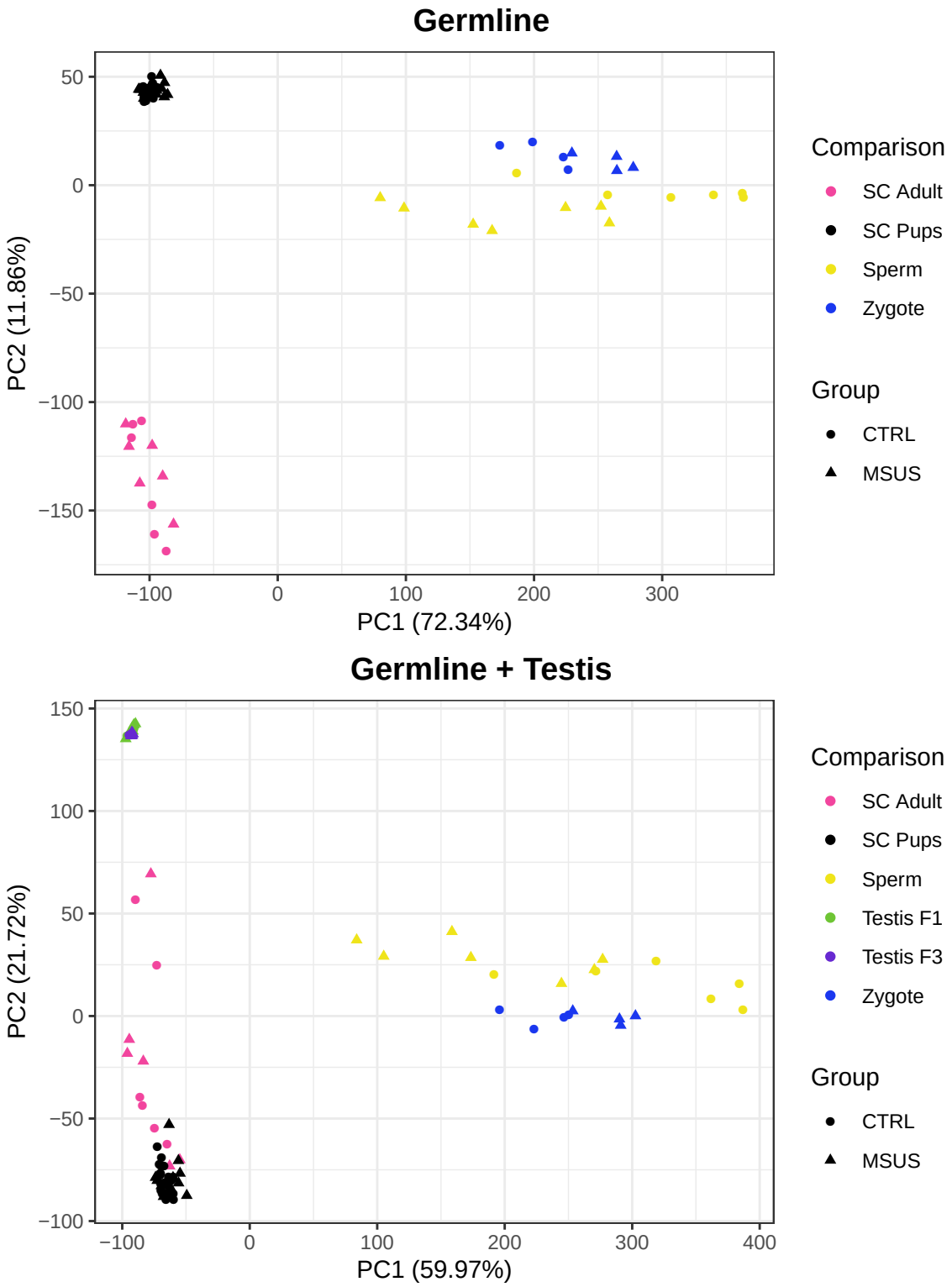

Figure 3.3: Principal Component Analysis - Germline. PCA of gene counts from germline and testis samples. Gene counts were obtained from the Salmon RNA-Seq quantification software. Lowly expressed genes were filtered and the remaining counts were normalized by $\log_2$CPM. Samples coming from the control group are plotted as circles, while samples coming from the MSUS group are plotted as triangles. The color of the points denotes which comparison they derive from. Top: Samples coming from germline tissues are plotted. Bottom: Samples coming from germline and testis tissues are plotted. Samples grouped together based on tissue type, but not generation or group.

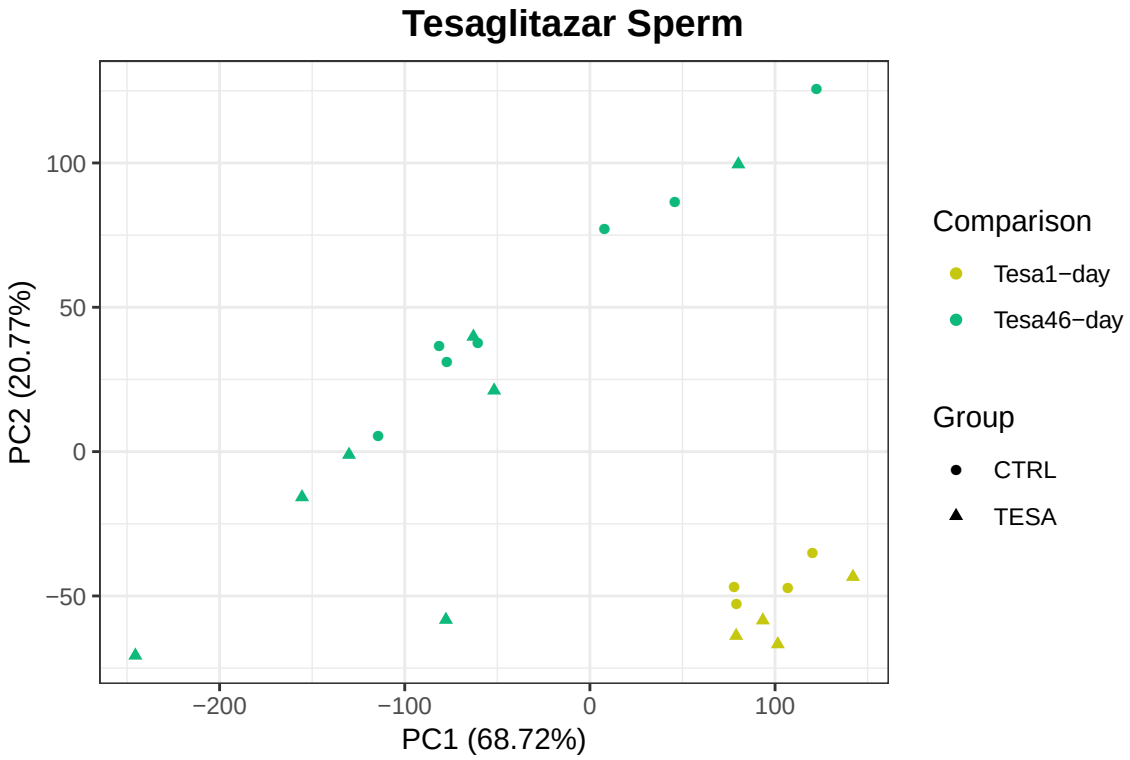

Figure 3.4: Principal Component Analysis - Tesaglitazar Sperm. PCA of gene counts from tesaglitazar sperm samples. Gene counts were obtained from the Salmon RNA-Seq quantification software. Lowly expressed genes were filtered and the remaining counts were normalized by $\log_2$CPM. Samples coming from the control group are plotted as circles, while samples coming from the MSUS group are plotted as triangles. The color of the points denotes which dataset they derive from.

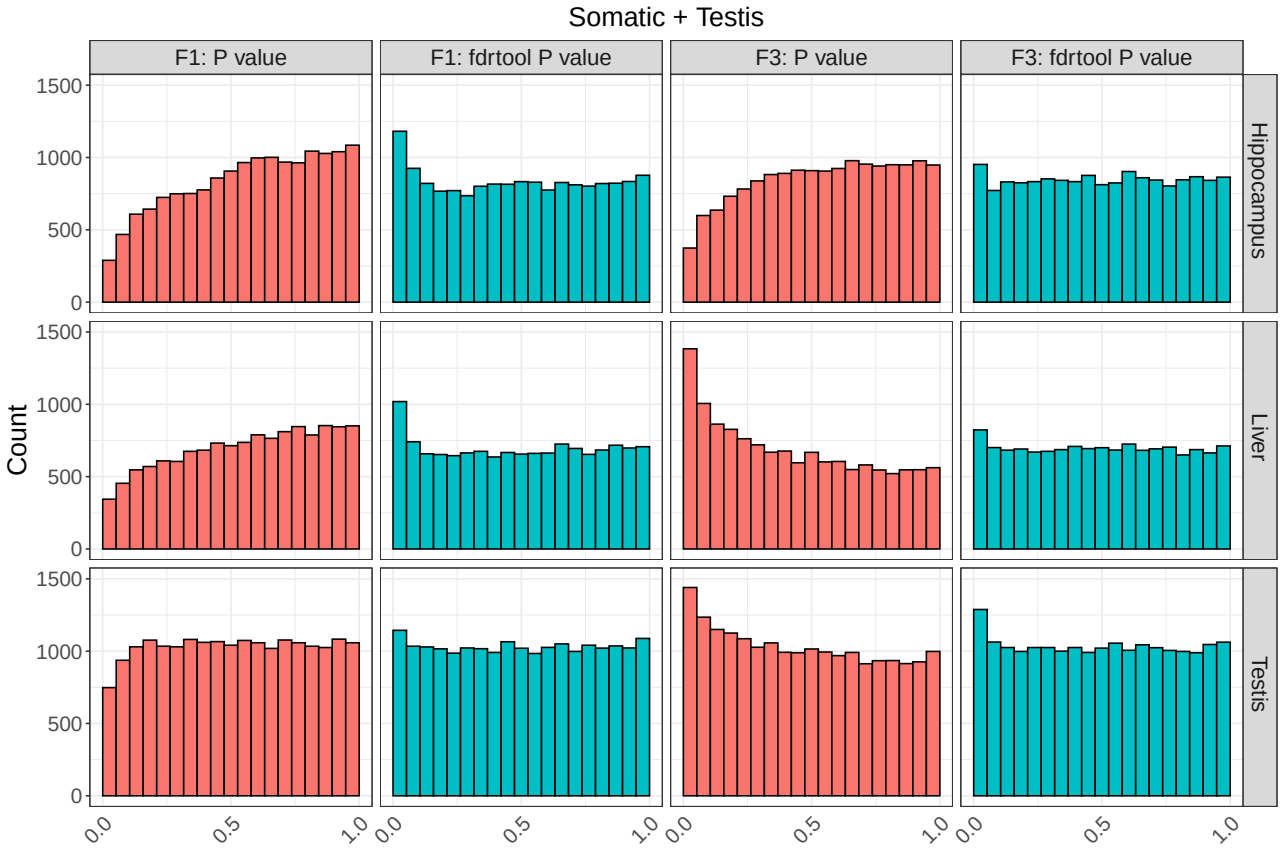

Figure 3.5: fdrtool P value Histogram - Somatic + Testis. Histograms of P values for F1 and F3 somatic and testis limma results before and after fdrtool correction. Histograms in red denote the original P value distribution, and histograms in blue denote the corrected distribution. Each row corresponds to a different tissue from the dataset.

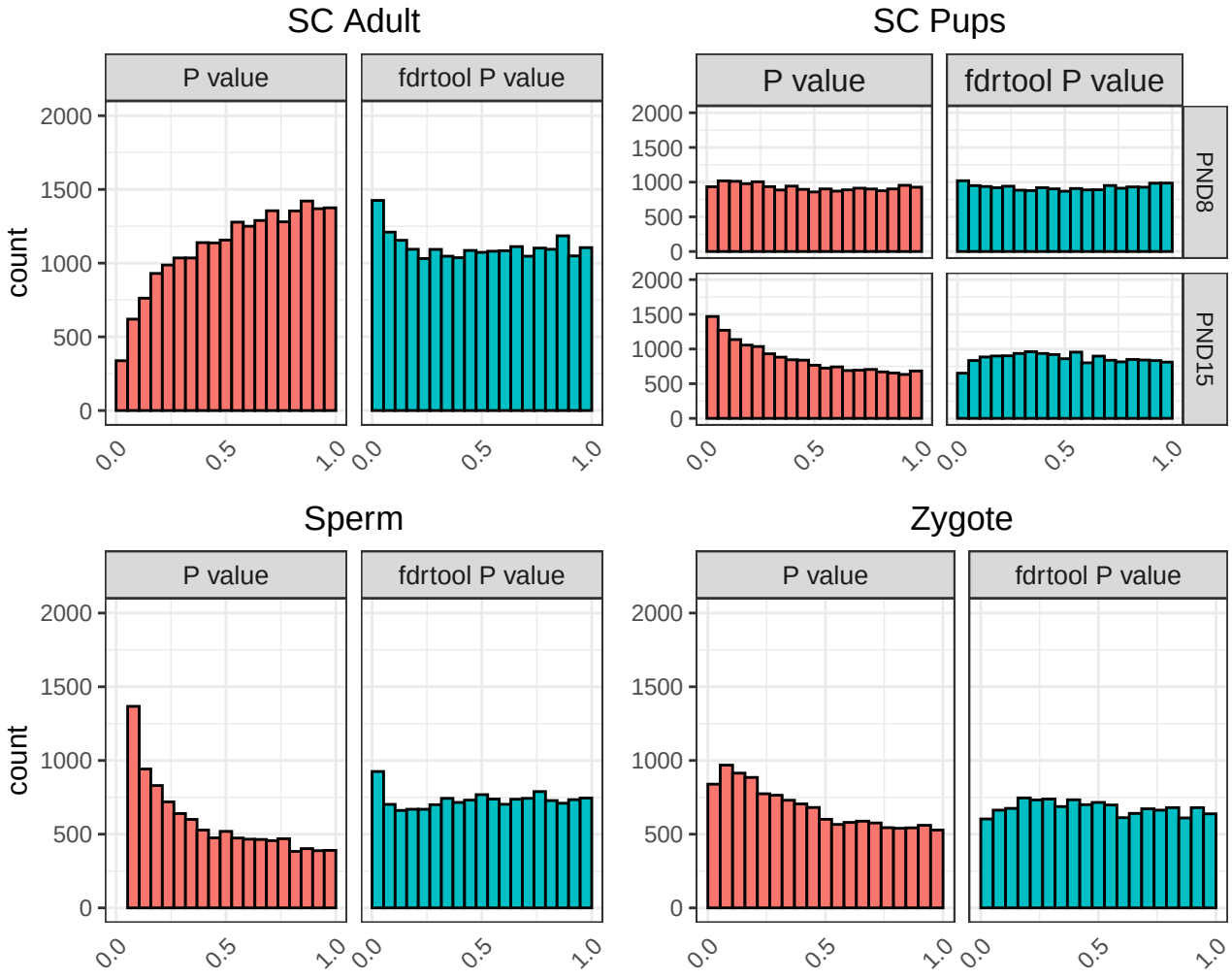

Figure 3.6: fdrtool P value Histogram - Germline. Histograms of P values for germline limma results before and after fdrtool correction. The histogram in red denotes the original P value distribution, and the histogram in blue denotes the corrected distribution.

more uniform distribution (Figure 3.6). For PND8, the corrected and uncorrected distributions are essentially the same. The original distribution of Sperm P values showed a high bias towards low P values and an uneven distribution, which was corrected by fdrtool. The fdrtool corrected Zygote P values also have a more uniform distribution. The lack of a clear peak near zero on the histograms for PND8, PND15, and Zygote is consistent with the lower number of differentially expressed genes found in these comparisons. The original distribution of P values for Tesa46-day

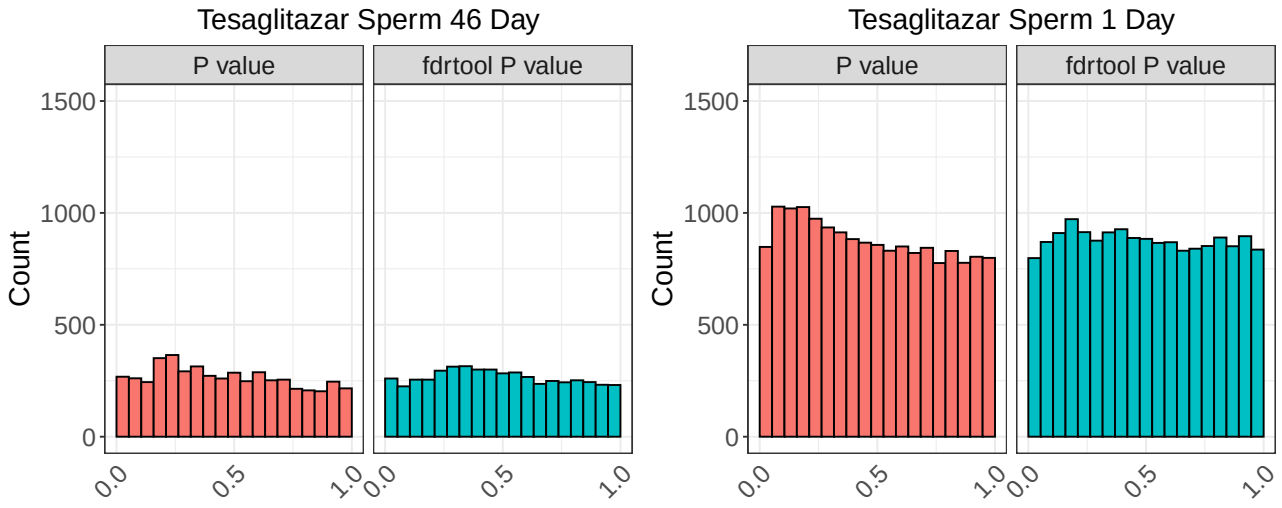

Figure 3.7: fdrtool P value Histogram - Tesaglitazar Sperm. Histograms of P values for Tesaglitazar Sperm limma results before and after fdrtool correction. The histograms in red denote the original P value distribution, and the histograms in blue denote the corrected distribution.

was slightly non-uniform, however the fdrtool corrected P values do not seem to be significantly more uniform (Figure 3.7). The fdrtool corrected P values for Tesa1-day have a moderately more uniform distribution. Unexpectedly, a peak near zero in both tesaglitazar histograms is not visible.

3.2.2 Proportion of up/down regulated differentially expressed genes

For all differential gene expression analyses, a stringent fdrtool P value of 0.005 was used to determine significance, rather than the Benjamini-Hochberg adjusted P values. Previous analyses done by Deepak Tanwar of the Mansuy laboratory used FDR correction, but no significant genes were found. Thus, a slightly relaxed criteria was used in the present analysis. Additionally, a $\log_2$ fold change threshold of 0.5 was applied, because expression changes less than that may not necessarily be biologically significant. It is important, given differential gene expression results, to see what proportion of genes in a given comparison are up-regulated and down-regulated. If there are clear patterns in these proportions, there may be underlying processes and mechanisms working to produce these results. For somatic and testis comparisons, there does not appear to be any distinct trend in either up or down-regulation of differentially expressed genes (Figure 3.8). An exception to this is the Liver F3 comparison; out of 70 differentially expressed genes, 66 were

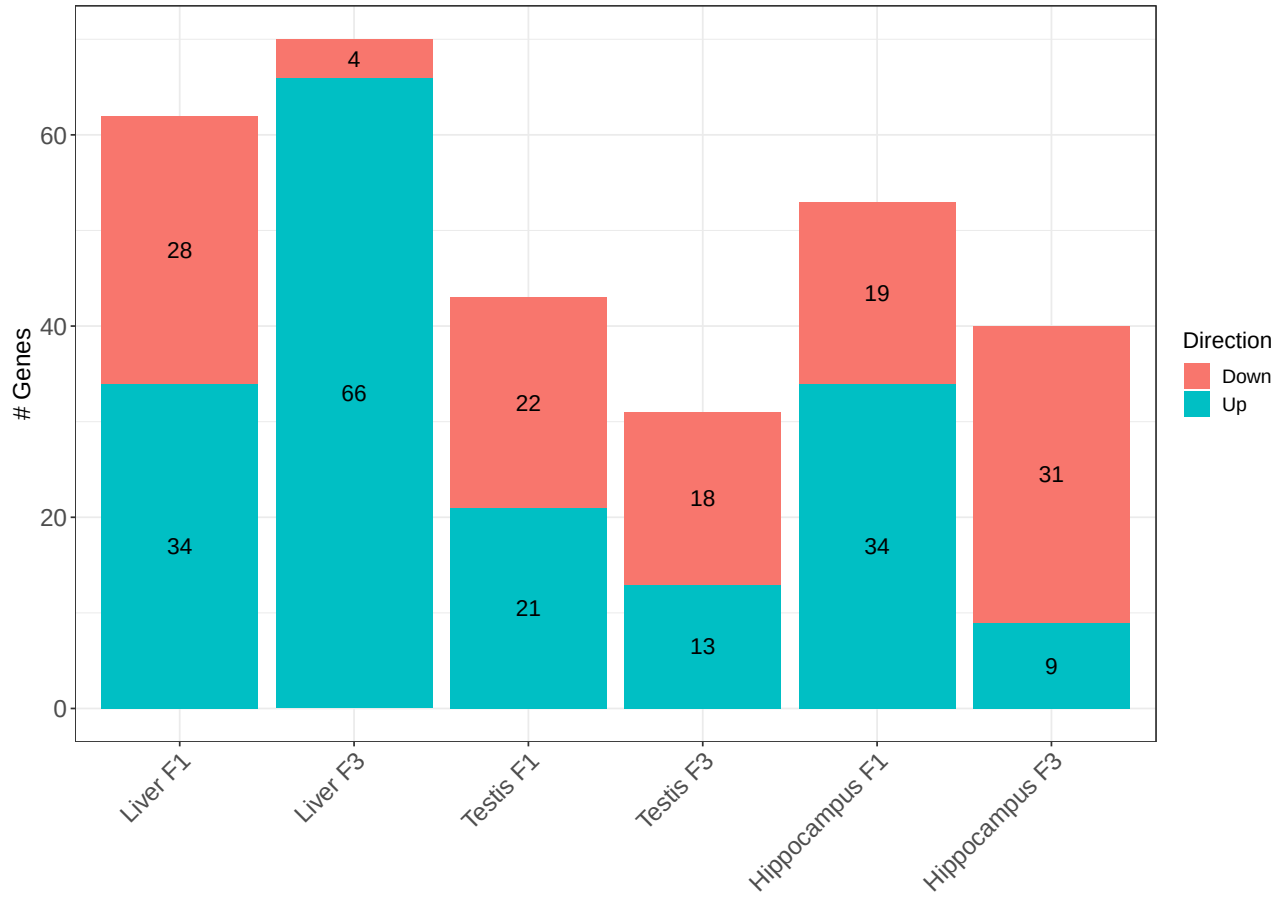

Figure 3.8: Differentially expressed genes - Somatic + testis (fdrtool P value $<$ 0.005; $\log_2$FC > 0.5). Barplot of the proportions of differentially expressed genes from somatic and testis limma results that are either up or down regulated. The genes were determined to be differentially expressed if both the absolute value of the $\log_2$ fold-change was greater than 0.5 and the fdrtool P value (not adjusted) was less than 0.005.

up-regulated (94%) in MSUS samples. Hippocampus F3 genes also had a bias towards down-regulated genes, with 31 out of 40 (78%) genes having a negative log fold change value in MSUS samples. As for the total number of significant differentially expressed genes, the numbers ranged from 31 in the Testis F3 comparison to 70 in the Liver F3 comparison. Between generations in each tissue, both testis and hippocampus showed a decrease in total genes, whereas there was an increase between F1 and F3 in liver samples. In germline comparisons, the proportions of up

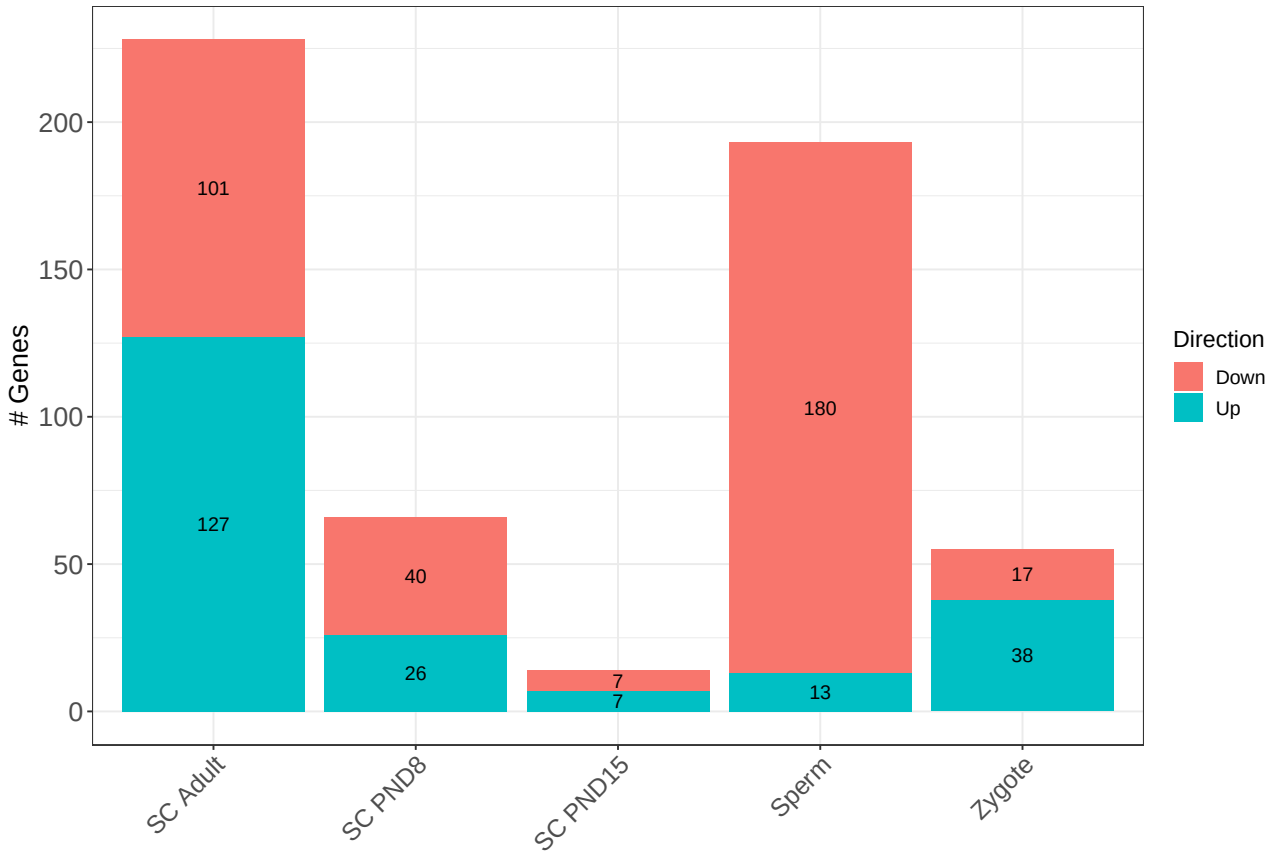

Figure 3.9: Differentially expressed genes - Germline (fdrtool P value < 0.005; $\log_2$FC > 0.5). Barplot of the proportions of differentially expressed genes from germline limma results that are either up or down regulated. The genes were determined to be differentially expressed if both the absolute value of the $\log_2$ fold-change was greater than 0.5 and the fdrtool P value (not adjusted) was less than 0.005.

and down regulated genes was mostly equal except for the Sperm comparison (Figure 3.9). Out of 193 Sperm DEGs, 180 (93%) were down-regulated. The number of differentially expressed genes varied considerably between comparisons. Both SC Adult and Sperm comparisons showed widespread differential gene expression, with SC Adult having 228 DEGs and Sperm having 193 DEGs. On the other hand, the SC pups had much lower levels of DEGs at PND8 (66 DEGs) and even less at PND15 (14 DEGs). When taken together with the SC Adult data, this presents a perplexing and uneven trend of decreasing DEGs between PND8 and PND15 followed by a substantial increase in DEGs in adult SC samples. The two tesaglitazar sperm comparisons show differences in both the number of total DEGs as well as the proportions of up and down regulated genes (Figure 3.10). The Tesa46-day comparison yielded 51 total DEGs, with 41 (82%)

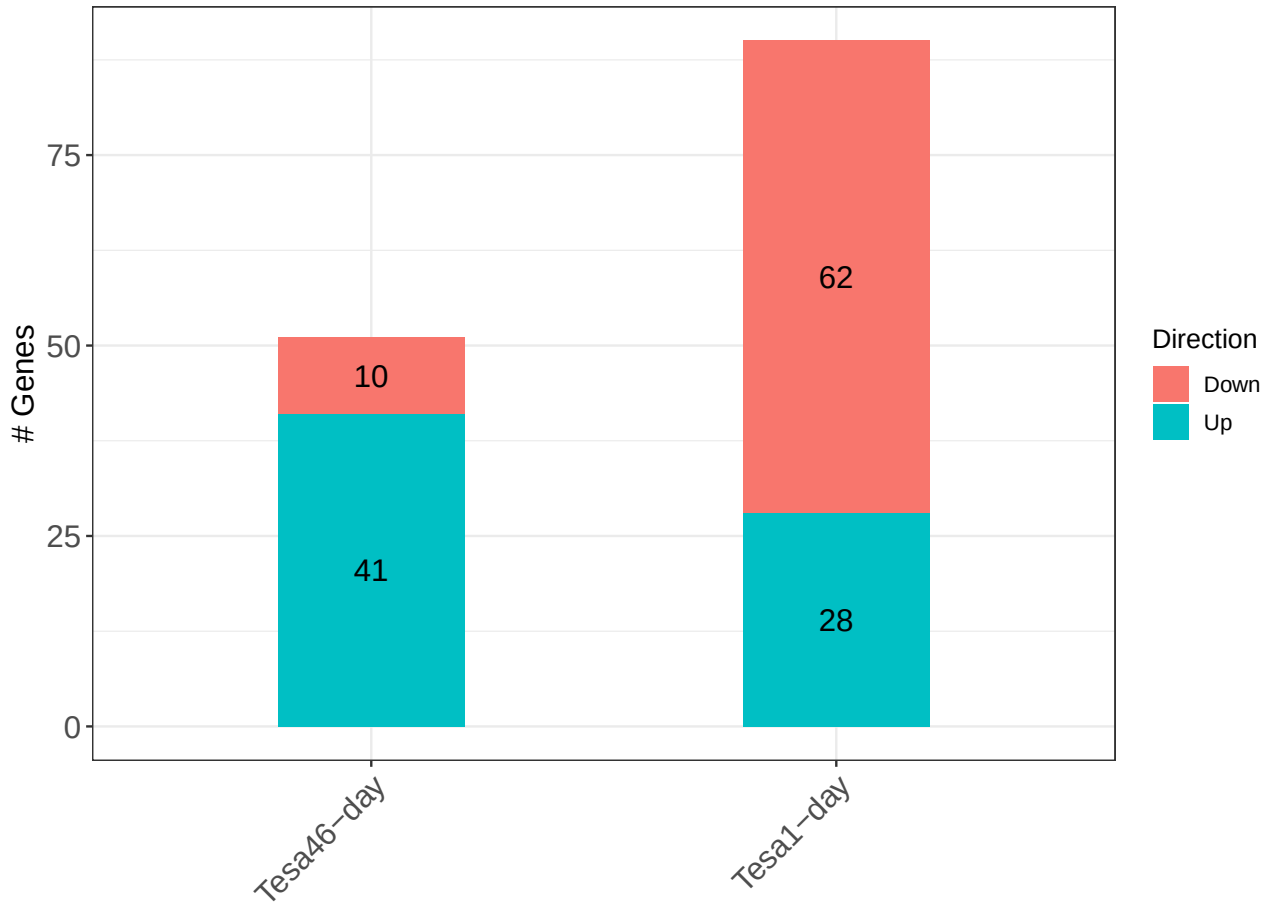

Figure 3.10: Differentially expressed genes - Tesaglitazar Sperm (fdrtool P value $<$ 0.005; $\log_2$FC > 0.5). Barplot of the proportions of differentially expressed genes from somatic and testis limma results that are either up or down regulated. The genes were determined to be differentially expressed if both the absolute value of the $\log_2$ fold-change was greater than 0.5 and the fdrtool P value (not adjusted) was less than 0.005.

being up-regulated. The Tesa1-day comparison, on the other hand, shows 90 total DEGs with 62 (69%) being down-regulated.

3.2.3 Top differentially expressed genes

In order to get a general idea of the types of genes being differentially expressed in MSUS mice, the top 5 DEGs from each comparison were taken and compiled into tables (Table 2, Table 3, Table 4). Certain results were ignored, specifically pseudogenes, ribosomal protein encoding genes, and uncharacterized/predicted genes. For all three groupings (somatic + testis, germline, tesaglitazar sperm), there appear to be genes involved in a variety of cellular functions and disease phenotypes, but it is unclear if any are directly related to the MSUS phenotype. However, some genes are involved or associated with functions and processes that have been found to be altered in MSUS mice, including metabolism, learning and memory, and epigenetic transcriptional regulation.

3.2.4 Intersections of differentially expressed genes

DEGs from each of the 13 comparisons were cross-checked with one another in order to see if there was any overlap of genes (Figure 3.11). Between the Zygote and Sperm DEGs, the genes Acss3 and Sfrp1 were present in both comparisons. The Sperm and SC Adult DEG lists both contained the two genes Ly6e and Glu1. The SC PND8 and Sperm DEG lists both contained Egf. The Liver F1 DEG list had intersections with three other DEG lists. With Tesa1-day, Liver F1 shares Lcn2, with Sperm it shares Cdh1, and with SC Adult it shares Ntrk2. In general it appears that the Sperm and Liver F1 comparisons have the most intersections with other datasets. However, it is evident that in general, there is little coherence between the differentially expressed genes found in each comparison. This suggests that there may not be any clear genetic signatures of the MSUS phenotype across tissue types.

3.2.5 Differentially expressed gene heatmaps

Heatmaps containing gene expression information were generated for each of the three groups of comparisons being analyzed, in order to assist in identifying gene signatures and general patterns of expression.. Within each group, the set of genes to be displayed in the heatmap were chosen by whether or not they met both the $\log_2$ fold change threshold (absolute value $>$ 0.5) and the fdrtool P value threshold (P value < 0.005) in the limma results of at least one of the comparisons. The union of genes for the somatic and testis comparisons yielded a total of 21 genes. The heatmap of these genes shows clear differential patterns of expression among the hippocampus, testis, and liver tissue samples (Figure 3.12). Within each tissue group, however, there does not appear to be any clear clustering of MSUS and control groups. Additionally, there is no clear clustering of F1 and F3 samples. The union of germline genes resulted in 97 genes. The samples clustered largely based on the dataset they originally derived from (Figure

Table 2: Top Differentially Expressed Genes - Somatic and Testis

Gene	Log_2FC	fdrtool P value	Comparison	Gene Information
Zfp708	-3.8977022	0.0000018	Liver F1	Nucleic acid binding
Socs2	-1.1671410	0.0006827	Liver F1	Low expression = weight increase
Ntrk2	-1.1294528	0.0010331	Liver F1	Learning and memory; CNS development
Lcn2	-1.0573156	0.0012755	Liver F1	Iron trafficking; innate immunity
Insl3	1.0473158	0.0000930	Liver F1	Testicular function
Obp2a	-1.5438857	0.0029992	Liver F3	Male fertility
Dclk2	1.5425571	0.0044997	Liver F3	Hippocampal organization
Galnt18	1.0572964	0.0002737	Liver F3	Oligosaccharide biosynthesis
Malat1	1.0532799	0.0004494	Liver F3	Transcriptional regulation
Kcnj10	1.0498671	0.0004053	Liver F3	Potassium buffering
Pagr1a	-2.9874646	0.0000621	Testis F1	Epigenetic transcriptional activation
Atp5g2	1.0373327	0.0004427	Testis F1	ATP synthase
Zfp871	-0.9442038	0.0009114	Testis F1	Nucleic acid binding
Ifit2	-0.8119816	0.0017805	Testis F1	Apoptosis; antiviral activity
Hist2h2aa2	0.6797890	0.0011891	Testis F1	Core histone
Hsd3b4	-3.6980292	0.0000314	Testis F3	Steroid reductase
Gm21887	-1.1004027	0.0007123	Testis F3	Predicted; Regulation of MAPK cascade
Slitrk2	-0.7926207	0.0014861	Testis F3	Synaptogenesis
Slc38a1	-0.6541612	0.0007181	Testis F3	Synthesis of glutamate and GABA
Adgrl4	-0.6416942	0.0014903	Testis F3	Regulates angiogenesis
Nutf2-ps1	-3.7845514	0.0000832	Hippocampus F1	Pseudogene
Pcdha8	-1.2454276	0.0010743	Hippocampus F1	Establishment of neuronal connections
Ccdc121	-1.1261680	0.0000004	Hippocampus F1	Contains coiled coil
Mfsd7a	0.9661871	0.0019554	Hippocampus F1	Transmembrane transport
Pappa2	0.9075232	0.0000339	Hippocampus F1	Regulator of insulin-ike growth factor
Lenep	-3.5792521	0.0000539	Hippocampus F3	Lens differentiation
Tgtp1	1.7416753	0.0000001	Hippocampus F3	Cell resistance to pathogens
Kremen2	1.2832816	0.0000000	Hippocampus F3	Inhibits Wnt/beta-catenin signaling
Ush1g	-1.1080043	0.0000001	Hippocampus F3	Required for normal hearing
Lrrc71	-0.8170867	0.0019783	Hippocampus F3	Unknown

Table 2: Top 5 differentially expressed genes from each somatic and testis limma analysis. Genes are ordered by the absolute value of their log_2 fold change and the comparison. Functions, roles, and diseases associated with each gene are listed.

Table 3: Top Differentially Expressed Genes - Germline

Gene	Log_2FC	fdrtool P value	Comparison	Gene Information
Urah	6.5490301	0.0000004	SC Adult	Hydrolase
Dnase1l1	5.5384874	0.0006581	SC Adult	DNase I like
Muc2	-5.5327471	0.0003283	SC Adult	Crohn disease; ulcerative colitis
Rragb	-5.5216899	0.0000528	SC Adult	Cellular response to amino acid availability
Clu	5.5091747	0.0046506	SC Adult	Stress induced apoptosis
Pramel7	-2.7932883	0.0008059	SC PND8	Represses DNA methylation in ESCs
Crygn	1.2357697	0.0048824	SC PND8	Function of auditory nuclei
Il20ra	1.1724732	0.0032689	SC PND8	IL-20 receptor
Cilp	1.1588467	0.0014221	SC PND8	Cartilage scaffolding
Cd59a	1.1295526	0.0024271	SC PND8	Inhibitor of the complement MAC action
Nlrp9b	1.2348474	0.0015128	SC PND15	Innate immunity
Trim5	1.1381279	0.0000599	SC PND15	Innate immunity
Slco4c1	1.0996640	0.0011220	SC PND15	Sperm maturation
Gjb3	-1.0081224	0.0036032	SC PND15	Gap junction protein
Sema6b	-0.6954708	0.0041220	SC PND15	CNS development
Dhcr7	-1.9649543	0.0000040	Sperm	Cholesterol production
Rn7s1	1.8409067	0.0006881	Sperm	Unknown
Galnt15	-1.8249503	0.0013315	Sperm	Oligosaccharide biosynthesis
Dleu2	-1.8226629	0.0049657	Sperm	miRNA host gene
Cd55	-1.7780161	0.0007203	Sperm	Regulates the complement cascade
H3f3c	4.5752668	0.0014407	Zygote	Histone
Khdc3	-4.1611500	0.0038939	Zygote	Proper spindle assembly
Gdf9	-2.6802858	0.0018532	Zygote	Gene expression regulation
Zp2	-2.5650639	0.0008999	Zygote	Secondary sperm receptor
Pparg	-1.7248238	0.0000585	Zygote	Regulates glucose homeostasis; obesity

Table 3: Top 5 differentially expressed genes from each germline limma analysis. Genes are ordered by the absolute value of their $\log_2$ fold change and the comparison. Functions, roles, and diseases associated with each gene are listed.

Table 4: Top Differentially Expressed Genes - Tesaglitazar

Gene	Log_2FC	fdrtool P value	Comparison	Gene Information
Rpl10	-2.058576	0.0036110	Tesa46-day	Embryonic brain development
Lyz2	-1.643588	0.0001804	Tesa46-day	Transglycosylation
Prg4	-1.499947	0.0015586	Tesa46-day	Lubrication of joints
Ppp1r7	1.478813	0.0021496	Tesa46-day	Mitotic cycle completion
Olfr1307	1.304443	0.0009497	Tesa46-day	Olfactory receptor
Fam177a	5.461946	0.0001212	Tesa1-day	Susceptibility to arthritis
Smok3c	-4.231941	0.0028464	Tesa1-day	Sperm motility kinase
Xkrx	-2.659886	0.0020318	Tesa1-day	Membrane transporter
Snora70	2.096844	0.0024430	Tesa1-day	snoRNA
Pstk	-1.976504	0.0009157	Tesa1-day	Selenocysteine biosythesis

Table 4: Top 5 differentially expressed genes from each Tesaglitazar sperm limma analysis. Genes are ordered by the absolute value of their $\log_2$ fold change and the comparison. Functions, roles, and diseases associated with each gene are listed.

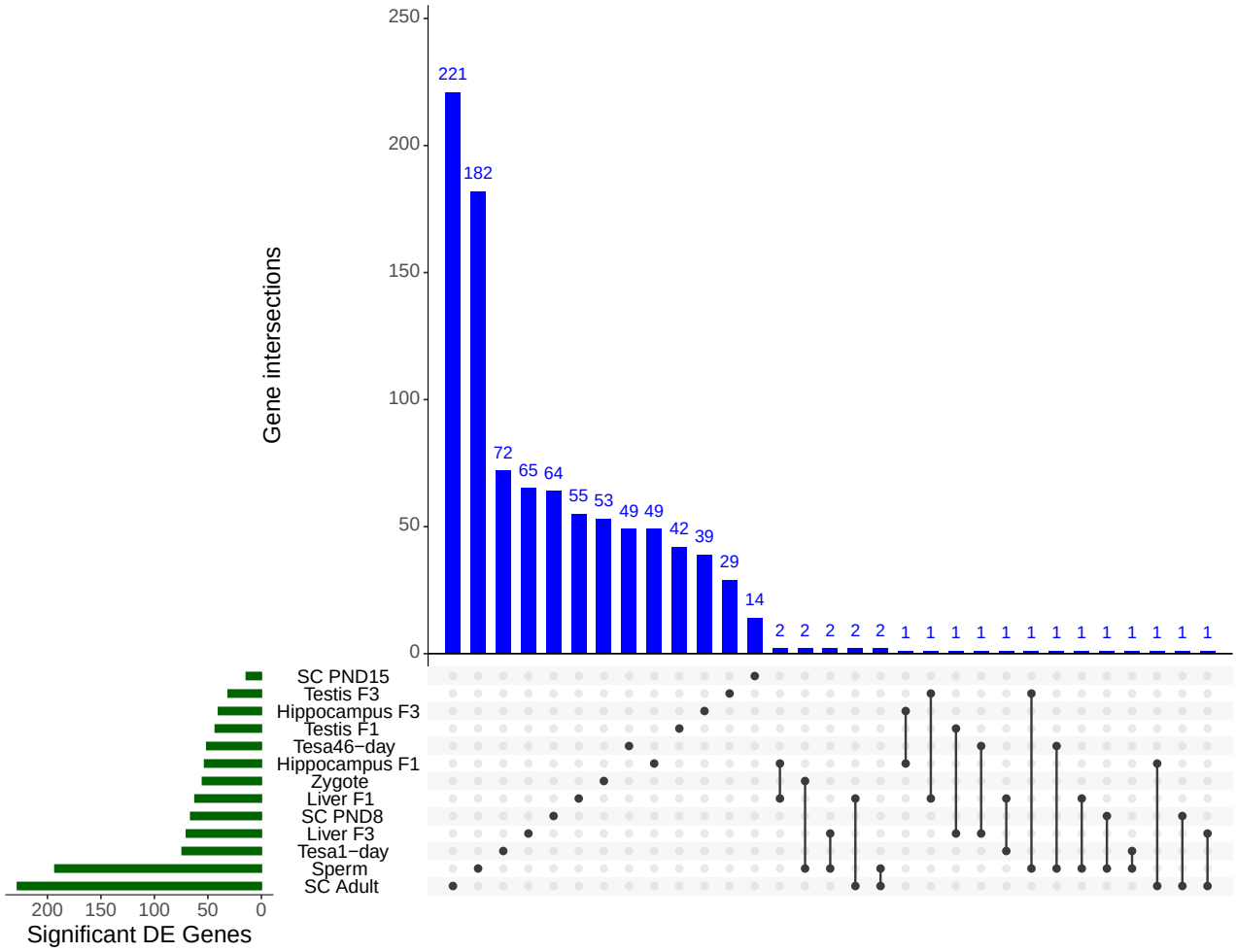

Figure 3.11: Upset plot of differential gene expression results. The 13 comparisons are listed alongside a barplot (in green) visualizing the amount of differentially expressed genes found in each comparison. Dots within the matrix signify intersections of genes, with the total number of genes in each intersection displayed in a barplot above the matrix (in blue). The intersections are ordered by frequency.

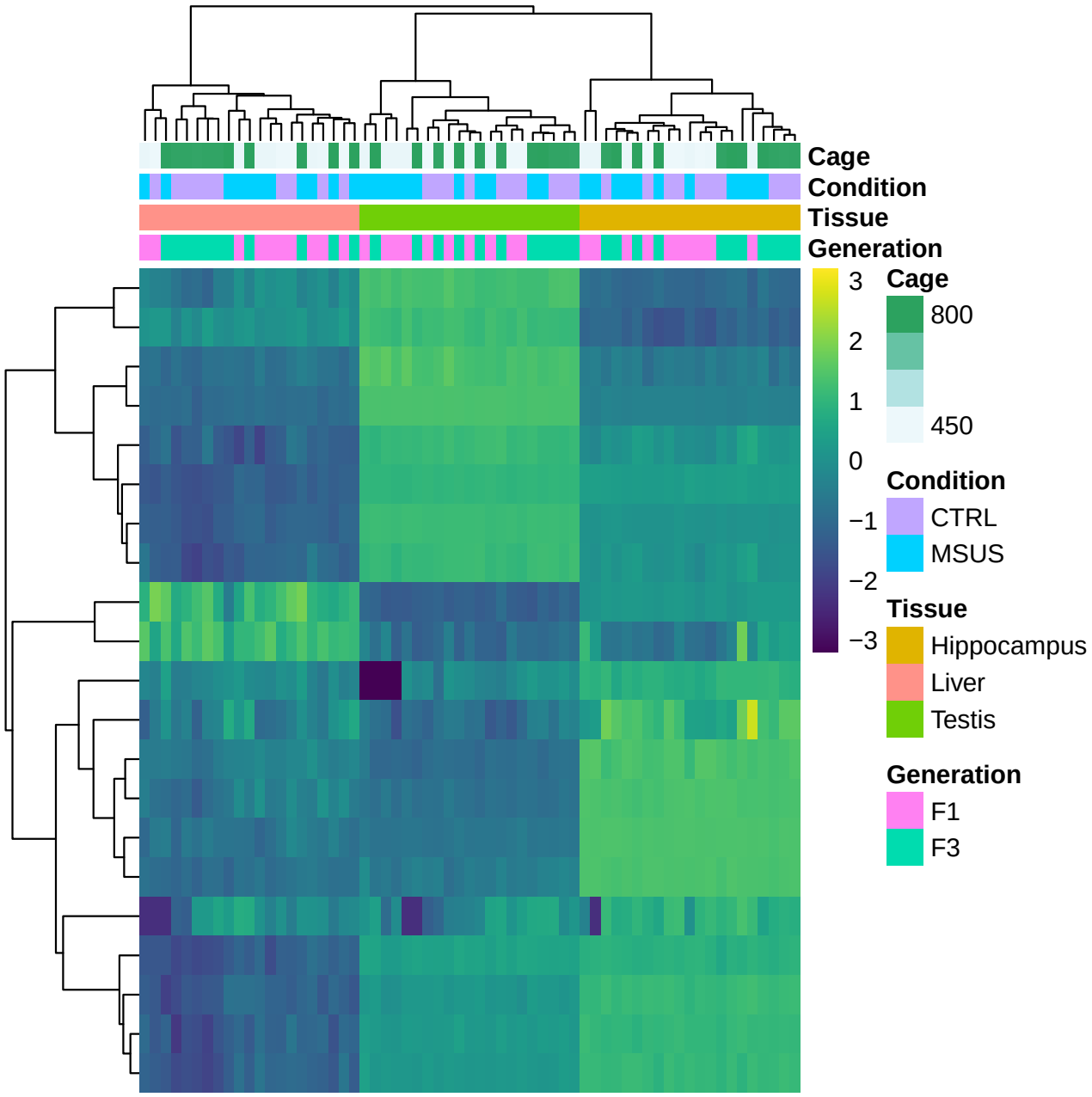

Figure 3.12: Heatmap - Somatic + testis (union). Heatmap of $\log_2$ Counts-per-million of gene counts from somatic and testis samples. The set of genes displayed are a result of a union of genes from each limma result that had a fdrtool P value of less than 0.005 and an absolute value of $\log_2$ fold change of greater than 0.5 in at least one of the limma results. A total of 21 genes met these criteria. Rows and columns were clustered. Samples are annotated with information including cage, condition, tissue, and generation.

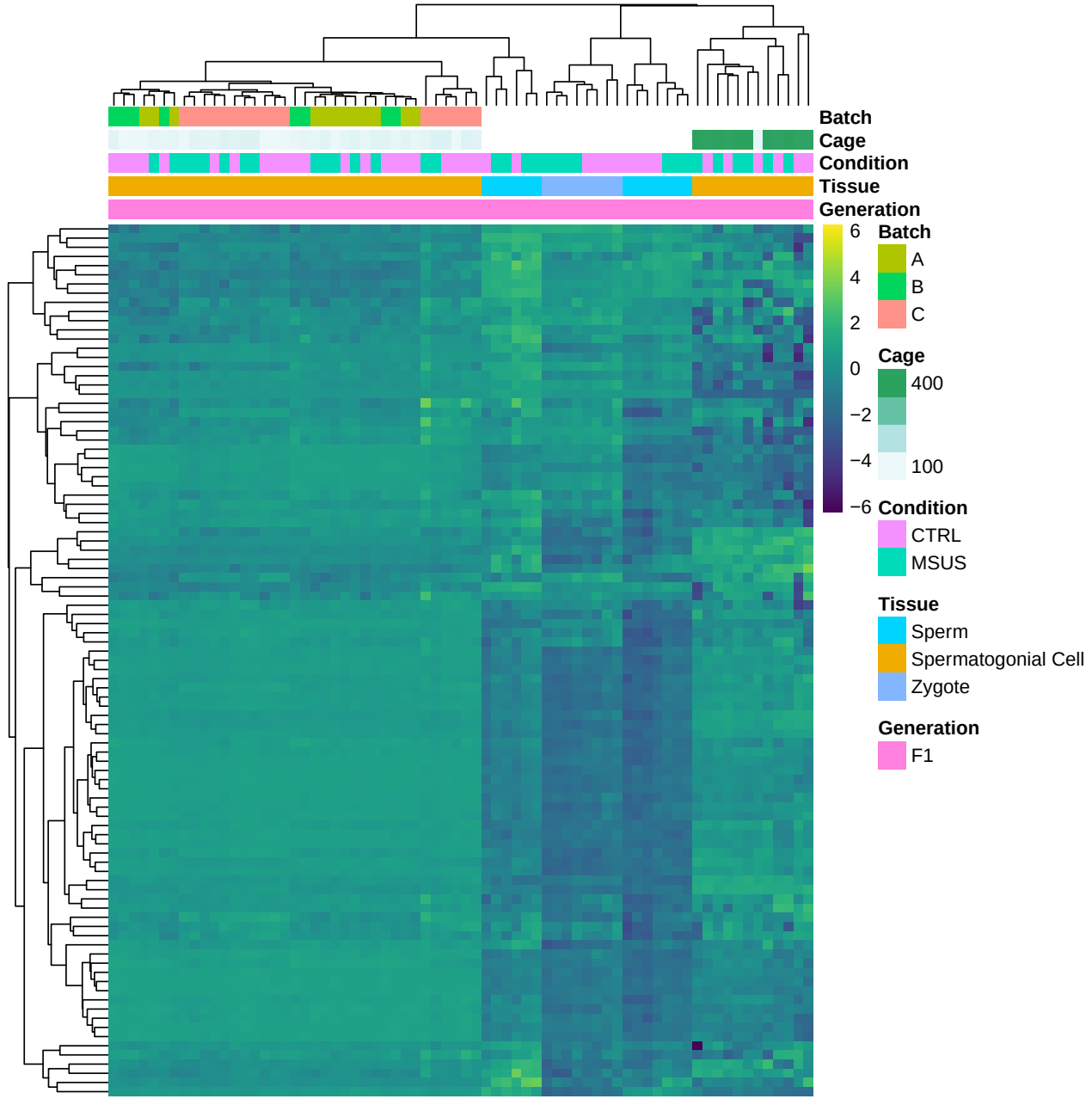

Figure 3.13: Heatmap - Germline (union). Heatmap of $\log_2$ Counts-per-million of gene counts from germline samples. The set of genes displayed are a result of a union of genes from each limma result that had a fdrtool P value of less than 0.005 and an absolute value of $\log_2$ fold change of greater than 0.5 in at least one of the limma results. A total of 97 genes met these criteria. Rows and columns were clustered. Samples are annotated with information including batch, cage, condition, tissue, and generation.

3.13). Within these 97 genes, most samples seem to have a $\log_2$CPM value of around 0, with some lower values as well. Unexpectedly, the SC pups samples did not cluster with the SC adult samples, and the Sperm samples split into two distinct clusters. Within each tissue group, there is no clustering of MSUS and control groups. Overall there appears to be about 5 discernable clusters with differing expression patterns among the 97 genes. Thirty genes passed the union criteria in the tesaglitazar sperm comparisons. There is no clear clustering of TESA and control samples, but there appears to be around three general clusters with differing expression patterns (Figure 3.14). Of particular note is the fifth sample from the left (Sperm-F1-TESA-KG8.2), which has an unusually low expression profile for all 30 genes, which is clearly different from all other samples.

3.2.6 Volcano plots

A common visualization method for DGE results is the volcano plot, which allows fast identification of DEGs and patterns of expression in large datasets.

In concordance with Figure 3.8, a volcano plot of somatic and testis genes shows that the Liver F3 comparison has a disproportionate amount of up-regulated genes, while the Hippocampus F3 comparison has a larger proportion of down-regulated genes (Figure 3.15). It is of note that there are many genes in each comparison that have a significant P value, but do not have a high enough fold change to be considered for this analysis. Similarly, the volcano plots for the Sperm and Zygote comparisons are in agreement with Figure 3.9, with the Sperm comparison DEGs being made up of mostly down-regulated genes and the Zygote comparison mostly up-regulated genes (Figure 3.16). Interestingly, there are no genes in either the SC Adult, Sperm, or Zygote comparisons that meet the P value threshold but not the fold change threshold. In particular, the SC Adult volcano plot has an unusual shape compared with the other volcano plots, in that it is much more spread out in a wing-like shape. However, this may be due in part to the larger number of genes represented on the SC Adult volcano plot (21,110 genes) compared to SC PND8 (17,600 genes), SC PND15 (16,424 genes), Sperm (13,910 genes), and Zygote (12,886 genes). The volcano plots for the tesaglitazar sperm comparisons corroborate the finding in Figure 3.10, showing a slight propensity for Tesa46-day genes to be up-regulated and for Tesa1-day genes to be down-regulated (Figure 3.17). Similar to the SC Adult, Sperm, and Zygote volcano plots, both tesaglitazar sperm volcano plots show no genes that meet the P value threshold but not the fold change threshold. An important observation is that the low gene-count filtering for the Tesa46-day comparison removed many more genes than in the other comparisons. While Tesa46-day has 5,042 genes plotted, Tesa1-day has 16,683 genes which passed the filtering step, which is closer in number to the germline comparisons.

3.2.7 Transcript level analysis

In order to assist with the experimental validation of laboratory results, all differential gene expression analyses were also completed at the transcript level, but the results of these analyses are not presented in this work. The transcript level results are beneficial to experimental

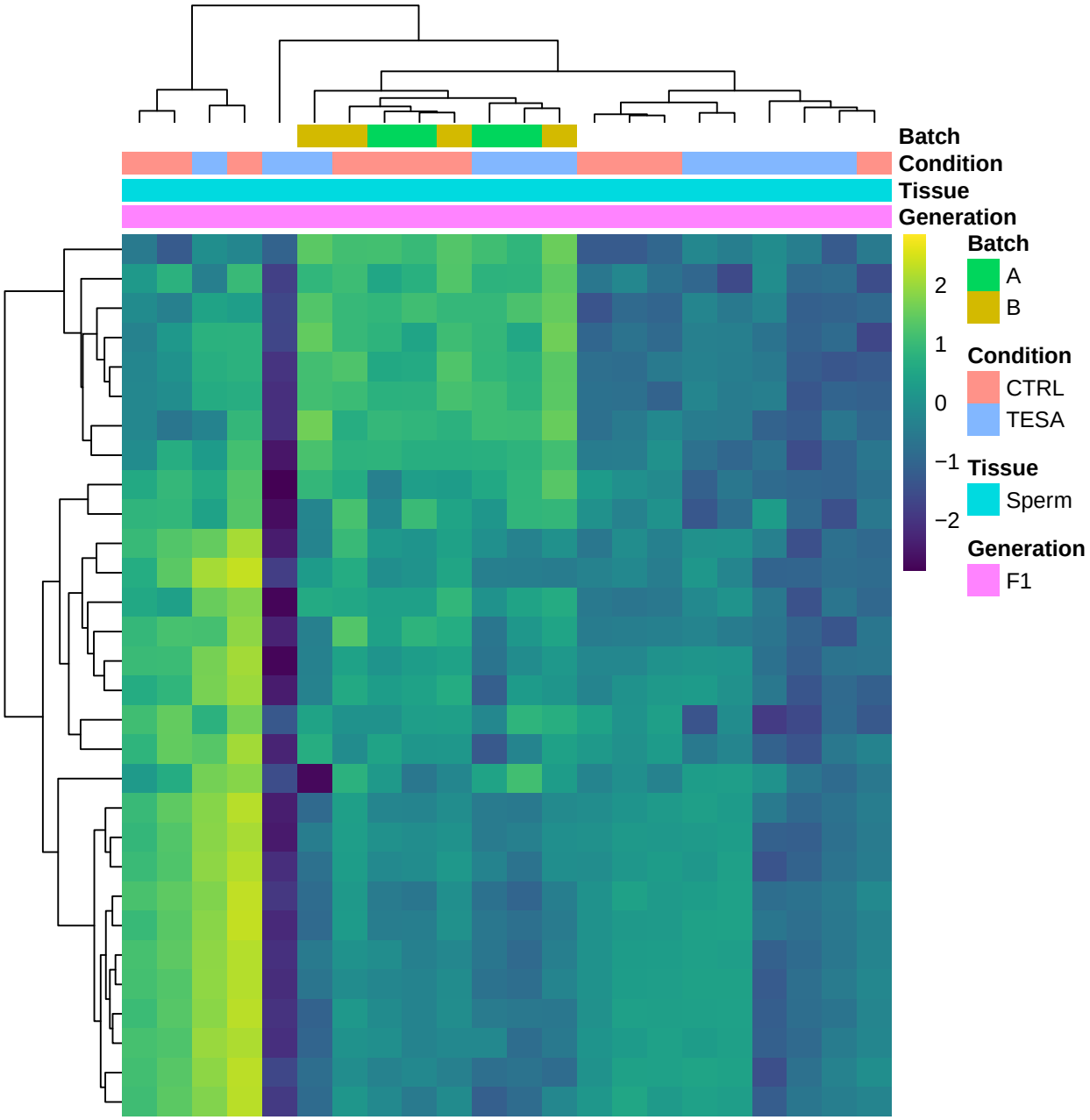

Figure 3.14: Heatmap - Tesaglitazar Sperm (union). Heatmap of $\log_2$ Counts-per-million of gene counts from Tesaglitazar sperm samples. The set of genes displayed are a result of a union of genes from each limma result that had a fdrtool P value of less than 0.005 and an absolute value of $\log_2$ fold change of greater than 0.5 in at least one of the limma results. A total of 30 genes met these criteria. Rows and columns were clustered. Samples are annotated with information including batch, condition, tissue, and generation.

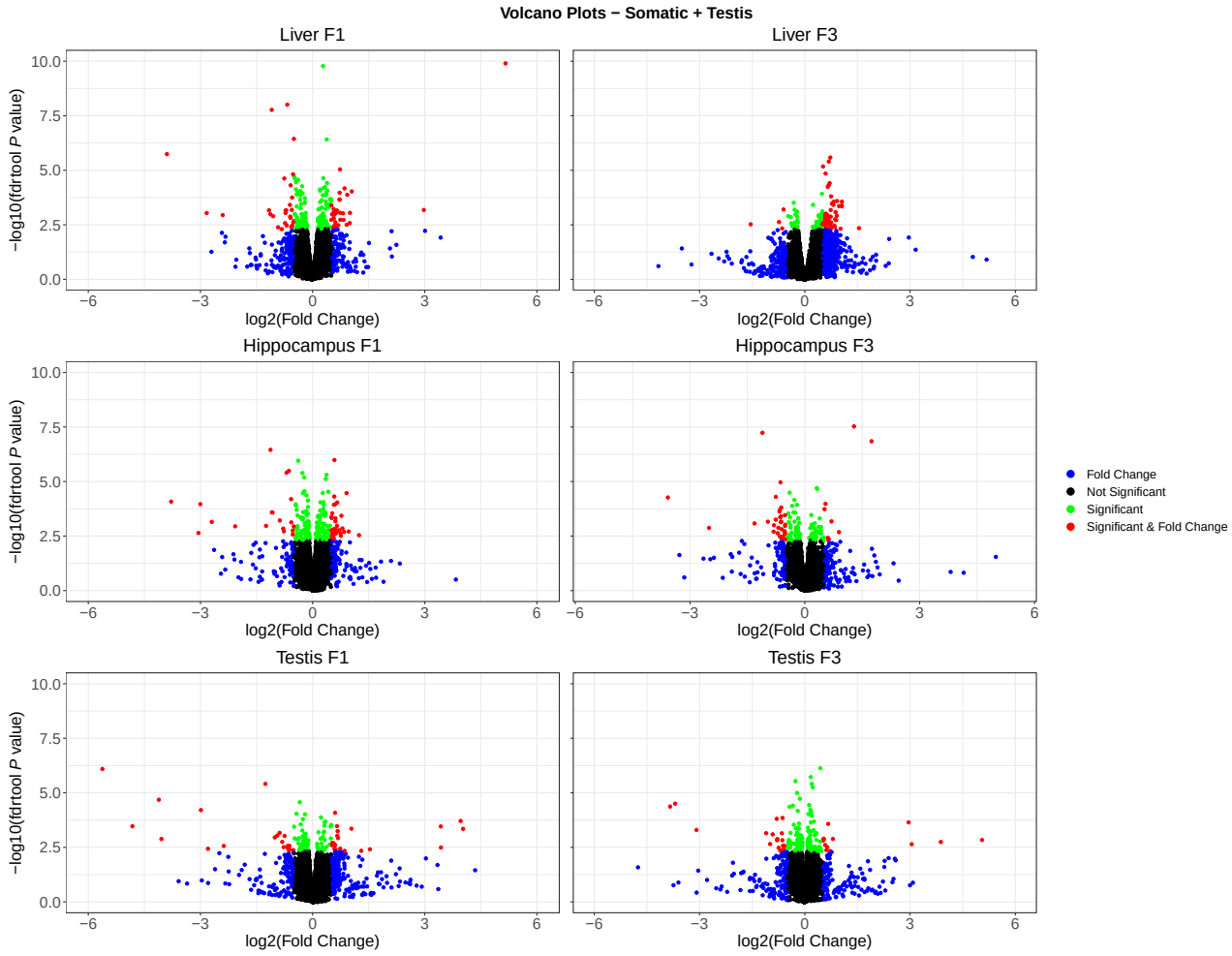

Figure 3.15: Volcano Plot - Somatic + Testis. Volcano plots for limma results of somatic and testis samples. The $\log_2$ fold change is plotted against the -$\log_{10}$ fdrtool P value. Black points represent genes that have an absolute $\log_2$ fold change of less than 0.5 and an fdrtool P value of greater than 0.005. Blue points represent genes that have an absolute $\log_2$ fold change of greater than 0.5, but an fdrtool P value of greater than 0.005. Green points represent genes that have an absolute $\log_2$ fold change of less than 0.5, but an fdrtool P value of less than 0.005. Red points represent genes that have an absolute $\log_2$ fold change of greater than 0.5 and an fdrtool P value of less than 0.005.

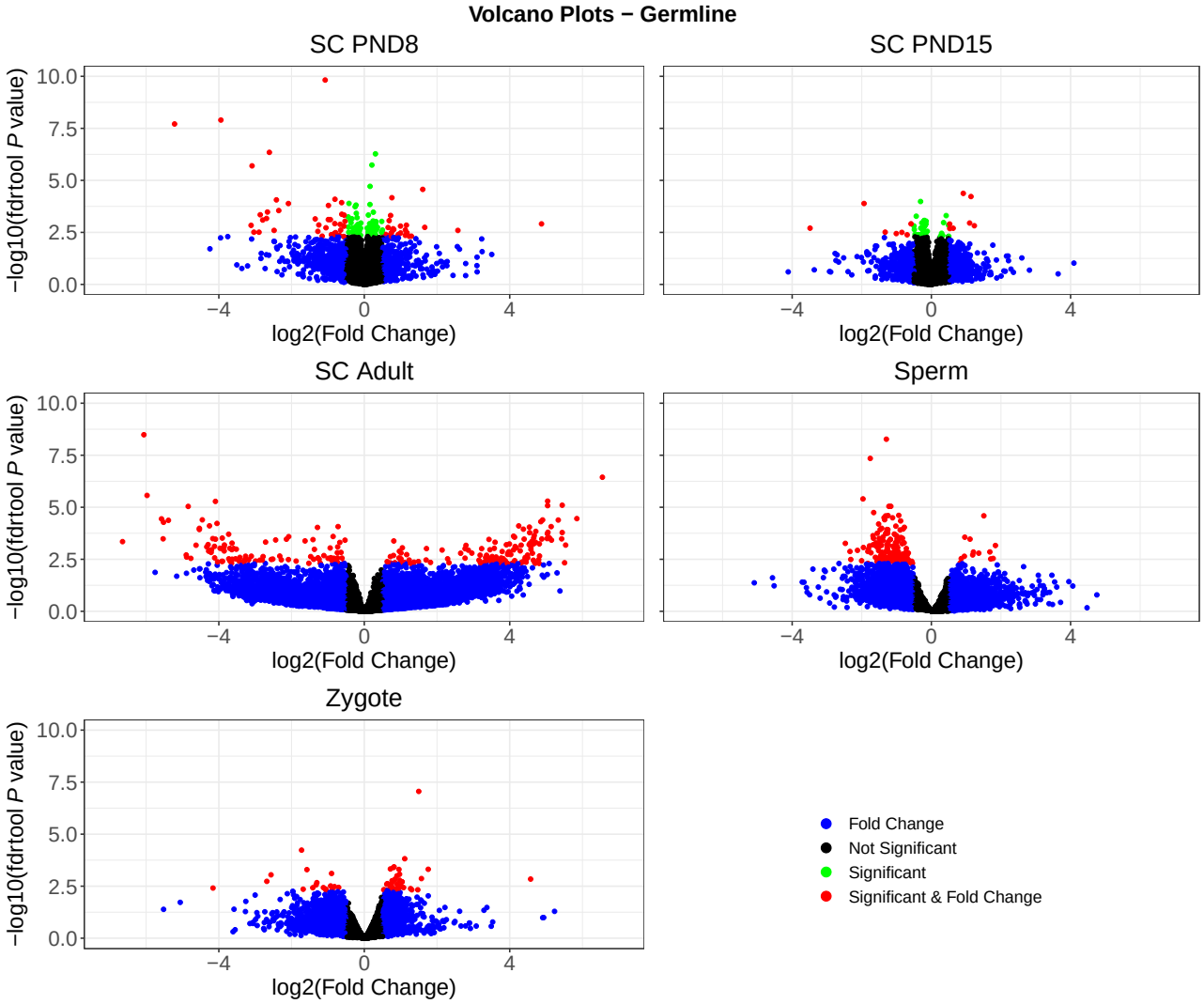

Figure 3.16: Volcano Plot - Germline. Volcano plots for limma results of germline samples. The $\log_2$ fold change is plotted against the -$\log_{10}$ fdrtool P value. Black points represent genes that have an absolute $\log_2$ fold change of less than 0.5 and an fdrtool P value of greater than 0.005. Blue points represent genes that have an absolute $\log_2$ fold change of greater than 0.5, but an fdrtool P value of greater than 0.005. Green points represent genes that have an absolute $\log_2$ fold change of less than 0.5, but an fdrtool P value of less than 0.005. Red points represent genes that have an absolute $\log_2$ fold change of greater than 0.5 and an fdrtool P value of less than 0.005.

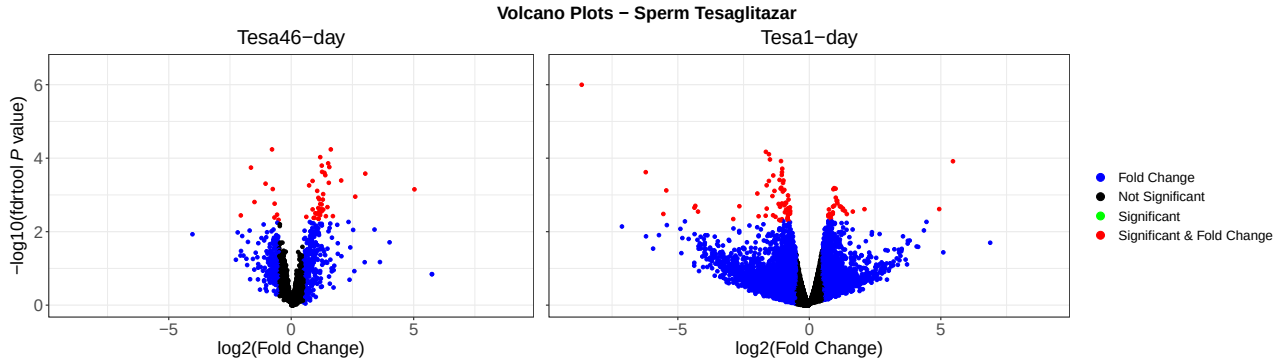

Figure 3.17: Volcano Plot - Tesaglitazar Sperm. Volcano plots for limma results of Tesaglitazar sperm samples. The $\log_2$ fold change is plotted against the -$\log_{10}$ fdrtool P value. Black points represent genes that have an absolute $\log_2$ fold change of less than 0.5 and an fdrtool P value of greater than 0.005. Blue points represent genes that have an absolute $\log_2$ fold change of greater than 0.5, but an fdrtool P value of greater than 0.005. Green points represent genes that have an absolute $\log_2$ fold change of less than 0.5, but an fdrtool P value of less than 0.005. Red points represent genes that have an absolute $\log_2$ fold change of greater than 0.5 and an fdrtool P value of less than 0.005.

biologists for designing primers based on specific isoforms of genes.

3.3 Pathway Analysis

3.3.1 Heatmaps of top KEGG pathways

The next step after performing differential gene expression analysis was to take those results and perform pathway analysis. This allows for a more comprehensive and biologically relevant look at the effects of the MSUS paradigm on the different tissues of the mice. Both KEGG and Reactome gene sets were used to do pathway analysis, but only the results from KEGG gene sets are reported in this work.

In 12 out of the 13 comparisons, with Tesa1-day being the one exception, the fGSEA analysis identified more significantly enriched pathways than CameraPR. This may be due to CameraPR having more stringent criteria in its algorithm, or more simply due to general differences in the algorithms. The Liver F3 and Sperm comparisons clearly yielded the highest numbers of significant pathways for both methods, with Liver F3 having 98 and 128 significant pathways from CameraPR and fGSEA analysis respectively, and Sperm having 104 and 143 significant pathways respectively. The first method used was CameraPR, a function that is part of the limma package. CameraPR reports a t-statistic that is used to determine the magnitude of the up or down regulation of each pathway. For each of the 13 MSUS (or TESA) versus control comparisons in this study, the t-statistics of the top 5 up-regulated KEGG pathways and top 5 down-regulated KEGG pathways were visualized on a heatmap (Figure 3.18). In certain cases, for example in Hippocampus F1, Testis F1, SC Adult, and SC PND8, there were no significant pathways reported. The pathways themselves were grouped by their higher-level categories

Figure 3.18: Heatmap - Top 5 and Bottom 5 Pathways - CameraPR. Heatmap of the t-statistics of the top 5 up and top 5 down regulated KEGG pathways from CameraPR analysis. Pathways are grouped and annotated based on their higher-level KEGG modules. Up regulated pathways are colored blue and down regulated pathways are colored red. The columns represent the 13 MSUS/TESA versus CTRL comparisons that were carried out on the 7 datasets, and are annotated based on which tissue type they belong to.

and the comparisons were grouped by the tissue-type they belonged to. The category "Disease" had the highest representation of unique pathways (11), but "Genetic Information Processing" had the highest amount of comparisons with significant pathways in that particular category (22). The pathway that was present in the most top 5 up or down regulated pathways was "Ribosome", which was in the top results in 6 out of the 13 comparisons. However, there was no general pattern to the regulation of the pathway, with 3 comparisons (Liver F1, Sperm, Tesa1-day) showing up-regulation of the "Ribosome" pathway and 3 comparisons (Hippocampus F3, Liver F3, Tesa46-day) showing down-regulation. The Tesa46-day and Tesa1-day comparisons showed considerable differences in their pathway regulation characteristics. Out of the 10 reported pathways for Tesa46-day, 9 were down-regulated, whereas for Tesa1-day, all 9 reported pathways were up-regulated. Interestingly, there were 4 pathways that overlapped between Sperm and Tesa1-day (also sperm samples), and all 4 pathways showed up-regulation of similar magnitude in both comparisons. Other significantly enriched pathways not shown on the heatmap from CameraPR analysis include "Thermogenesis" (Tesa46-day - down-regulated; Sperm - up-regulated; SC PND15 - down-regulated). The second method to perform pathway analysis was fGSEA, a standalone R package that is based on the popular GSEA method. Rather than report a t-statistic for each pathway, fGSEA reports a normalized enrichment score (NES) that is used to determine the magnitude of up or down regulation of a pathway. Analogously to the CameraPR analysis, for each of the 13 MSUS (or TESA) versus control comparisons in this study, the NES of the top 5 up-regulated KEGG pathways and top 5 down-regulated KEGG pathways were visualized on a heatmap (Figure 3.19). Liver F1, Testis F1, SC Adult, Tesa1-day all had no significantly enriched pathways reported. Once again, the pathways themselves were grouped by their higher-level categories and the comparisons were grouped by the tissue-type they belonged to. Similarly to the CameraPR results, "Disease" was the most represented higher-level KEGG category, with 14 pathways present on the heatmap. "Genetic Information Processing" had the highest amount of comparisons with significant pathways in that category (22). "Oxidative phosphorylation" and "Ribosome" were the pathways with the most representation in the 13 comparisons, both of which being present in 5 out of the 13. Other significantly enriched pathways not shown on the heatmap from fGSEA analysis include "PPAR signalling pathway" (Zygote - down-regulated), "Fatty acid degradation" (Zygote - down-regulated), "Thermogenesis" (SC PND15 - down-regulated; Tesa46-day - down-regulated), and "Metabolic pathways" (Sperm - down-regulated; SC PND15 - up-regulated; Testis F3 - down-regulated). These 4 pathways are parts of biological functions and processes that have been shown to be altered in the MSUS phenotype.

Dot plots showing the top pathway results for each individual comparison for both CameraPR and fGSEA analysis can be found in Appendix A.

3.4 Differential Exon Usage

The final step in this work was to perform differential exon usage (DEU) analysis, in order to get a more clear picture of how alternative splicing is affected in MSUS and to see which specific genes and exons are targeted by this process. Of important note is that for all DEU analysis, a Benjamini-Hochberg multiple testing adjusted P value was calculated and used to determine significant exons, rather than the fdrtool P value used in previous DGE analysis. Additionally,

Figure 3.19: Heatmap - Top 5 and Bottom 5 Pathways - fGSEA. Heatmap of the normalized enrichment score (NES) of the top 5 up and top 5 down regulated KEGG pathways from fGSEA analysis. Pathways are grouped and annotated based on their higher-level KEGG modules. Up regulated pathways are colored blue and down regulated pathways are colored red. The columns represent the 13 MSUS/TESA versus CTRL comparisons that were carried out on the 7 datasets, and are annotated based on which tissue type they belong to.

no fold change cutoff was set for DEU results.

3.4.1 MA plots

One of the simplest and easy to interpret representations of DEXSeq results is the MA plot, which visualizes the relationship between the expression and fold change of each tested exon. MA plots were generated from somatic and testis DEXSeq results, with an FDR of 0.1 chosen to determine significance (Figure 3.20). Strikingly, there was no differential exon usage in any of the six somatic or testis comparisons. With an FDR of 0.1, MA plots were also generated for germline DEXSeq results (Figure 3.21). In contrast to the somatic and testis results, there was differential exon usage in the SC Adult, SC PND8, SC PND15, and Sperm comparisons. In particular, SC PND8 appears to have extensive DEU, with 3,280 total exons showing significant DEU. With a total of 664,251 exons tested in SC PND8, this represents 0.49% of the total exons. The Sperm comparison yielded the second highest count of significant exons with a total of 578 passing multiple testing correction. SC PND15 had 214, and SC Adult had 69. The Zygote comparison yielded no differential exon usage.

Several genes which have been previously mentioned in the literature as possibly related to the MSUS phenotype had exons which passed multiple testing correction. In SC PND8, there was DEU in 4 MSUS-related genes; Tsc1, Tsc2, Ctnnb1, and Slc1a2. These genes had DEU in 10, 12, 4, and 2 exons with differential usage, respectively. In Sperm, Ctnnb1 had 1 significant exon usage change. MA plots for tesaglitazar sperm DEXSeq results were created, also with an FDR of 0.1 (Figure 3.22). There is widespread DEU in the Tesa1-day comparison, totalling 13,090 significant exons. In Tesa46-day, there appears to be no differential exon usage. However, although there are no significant results, there is a clear trend in the MA plot for Tesa46-day, with a disproportionate amount of the exons showing a negative $\log_2$ fold change. This observation is not present in any of the other 12 comparison MA plots. Three MSUS-associated genes from Tesa1-day had exons which passed multiple testing correction. Dnmt3a had DEU in 1 exon, Tsc1 had DEU in 18 exons, and Tsc2 had DEU in 8 exons.

3.4.2 Top DEXSeq results

To get a general idea of the types of genes displaying differential exon usage in MSUS mice, the top 5 DEU results from each comparison where there was DEU present were taken and compiled into a table (Table 5). There appears to be DEU in genes involved in a variety of cellular functions and disease phenotypes, but it is unclear if any are directly related to the MSUS phenotype. However some genes are involved or associated with functions and processes that have been found to be altered in MSUS mice. It is also unclear as to the effect that the differential usage of a particular exon in each gene would have on the functional description given in the table.

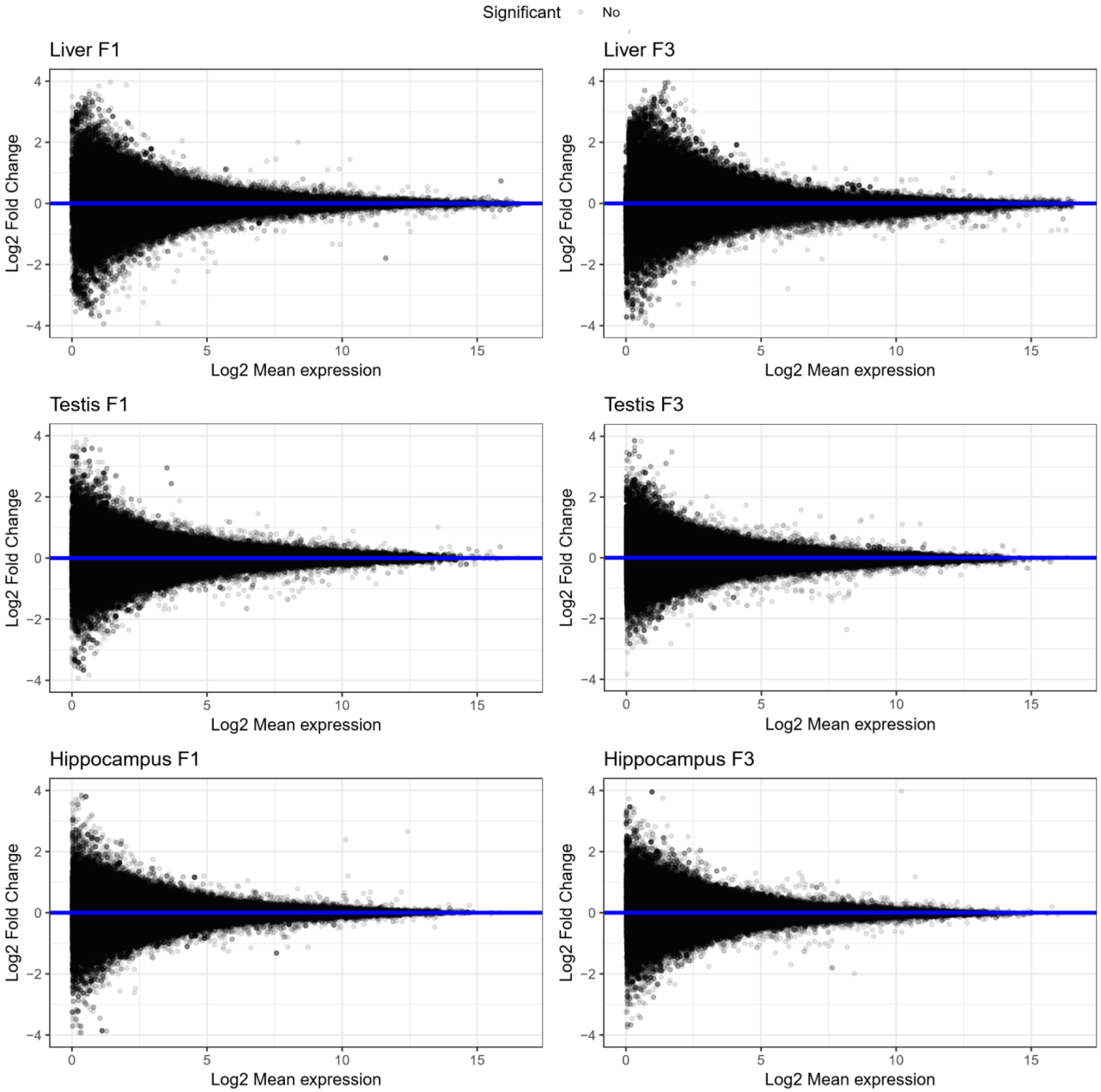

Figure 3.20: DEXSeq MA Plots - Somatic + Testis. MA plot of exons from somatic and testis samples. Data was obtained during DEXSeq analysis. FDR was set to 0.1. Exons that were determined to be significantly differentially used are colored in red. The x-axis is the $\log_2$ mean expression value, while the y-axis is the $\log_2$ fold change.

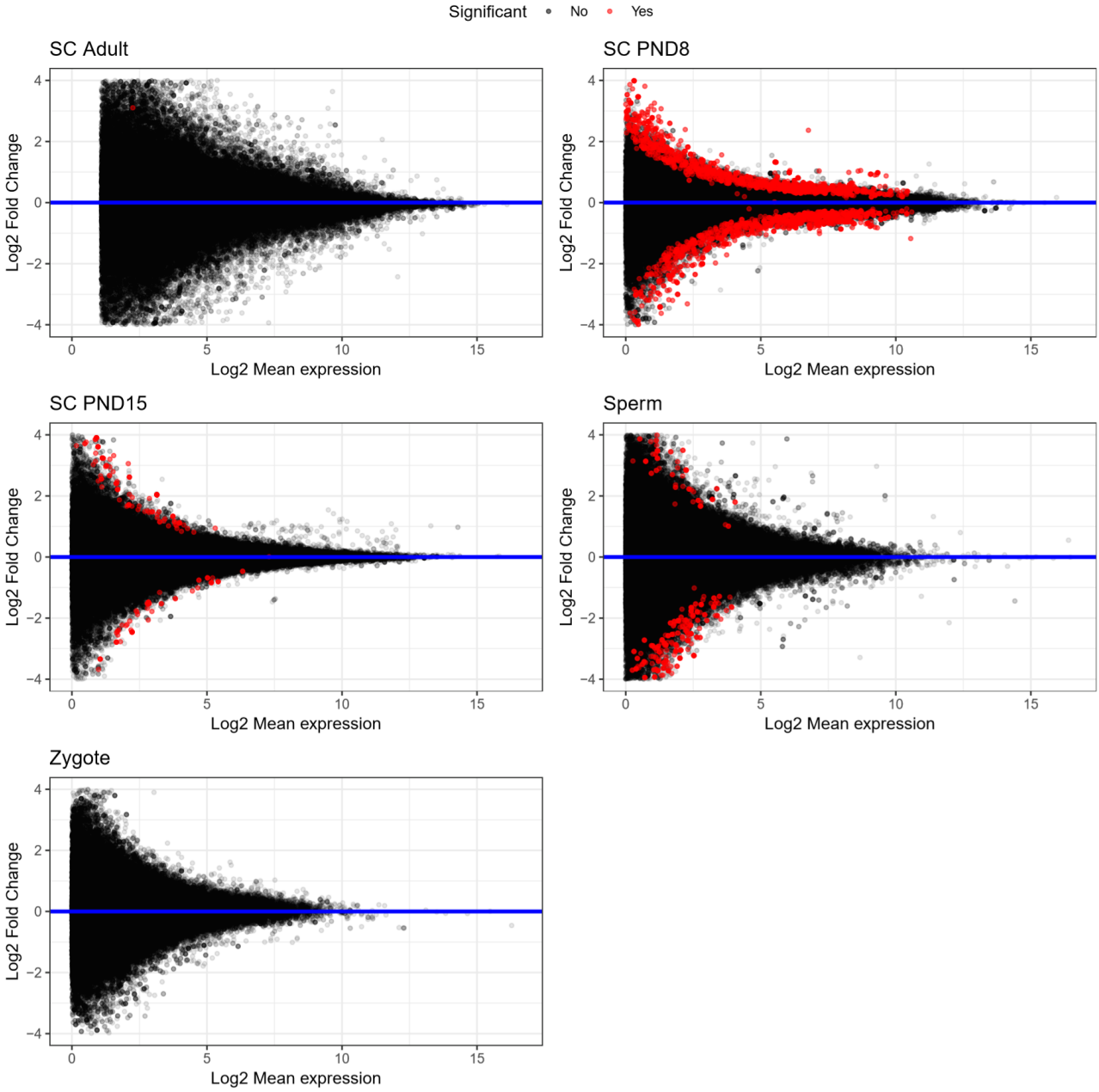

Figure 3.21: DEXSeq MA Plots - Germline. MA plot of exons from germline samples. Data was obtained during DEXSeq analysis. FDR was set to 0.1. Exons that were determined to be significantly differentially used are colored in red. The x-axis is the $\log_2$ mean expression value, while the y-axis is the $\log_2$ fold change.

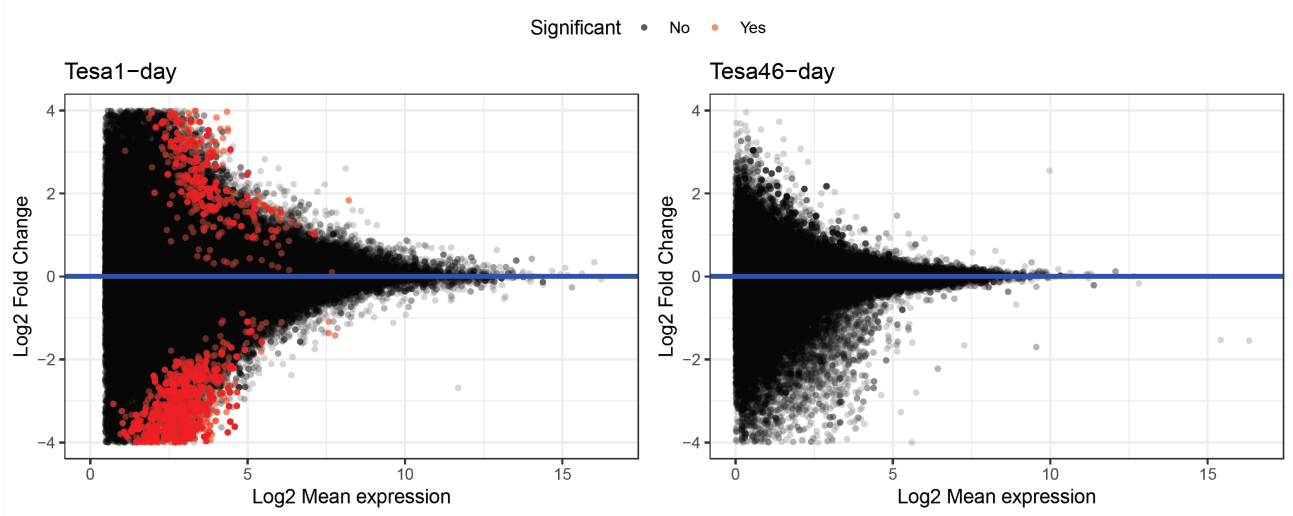

Figure 3.22: DEXSeq MA Plots - Tesaglitazar Sperm. MA plot of exons from Tesaglitazar sperm samples. Data was obtained during DEXSeq analysis. FDR was set to 0.1. Exons that were determined to be significantly differentially used are colored in red. The x-axis is the $\log_2$ mean expression value, while the y-axis is the $\log_2$ fold change.

Table 5: Top DEXSeq Results

Gene	Exon	Log_2FC	Adj. P value	Comparison	Gene Information
Grb10	ENSMUSG00000020176.17.E69	19.85	0.0486823	SC Adult	Interaction w/ insulin receptors
Kif16b	ENSMUSG00000038844.10.E12	-18.23	0.0486823	SC Adult	Intracellular trafficking
Adck3	ENSMUSG00000026489.13.E64	15.49	0.0486823	SC Adult	DNA damage response
Eif2ak4	ENSMUSG00000005102.13.E123	5.37	0.0486823	SC Adult	Protein synthesis downregulation
Oser1	ENSMUSG00000035399.12.E19	3.11	0.0486823	SC Adult	Unknown
Pex3	ENSMUSG00000019809.16.E1	4.82	0.0155460	SC PND15	Peroxisome biosythesis
Larp1b	ENSMUSG00000025762.14.E57	3.61	0.0155460	SC PND15	RNA binding
Rapgef5	ENSMUSG00000041992.9.E16	3.32	0.0099639	SC PND15	Signal transduction
Cic	ENSMUSG00000005442.13.E5	-2.77	0.0155460	SC PND15	Transcription repression
Atrn	ENSMUSG00000027312.14.E59	1.92	0.0035632	SC PND15	Energy homeostasis
Chchd10	ENSMUSG00000049422.7.E1	15.67	0.0000006	SC PND8	Dementia in humans
Slc10a7	ENSMUSG00000031684.11.E38	3.04	0.0001714	SC PND8	Skeletal development
Tsc22d1	ENSMUSG00000022010.19.E26	2.68	0.0001714	SC PND8	Tumor suppression
Cobll1	ENSMUSG00000034903.18.E98	2.65	0.0001775	SC PND8	Lower Insulin resistance
Pole	ENSMUSG00000007080.14.E71	0.46	0.0003305	SC PND8	Growth disorders; DNA repair
Cbfa2t2	ENSMUSG00000038533.15.E10	-17.27	0.0019718	Sperm	Osteogenic differentiation
Ube2o	ENSMUSG00000020802.8.E28	-4.55	0.0019718	Sperm	Ubiquitin conjugation
Vps13d	ENSMUSG00000020220.16.E119	-3.39	0.0025344	Sperm	IL-6 production
Fryl	ENSMUSG00000070733.13.E192	-2.66	0.0036680	Sperm	Transcription coactivator
Sec23ip	ENSMUSG00000055319.8.E5	-1.75	0.0019718	Sperm	ER-Golgi transport
Chd7	ENSMUSG00000041235.12.E50	-17.70	0.0000000	Tesa1-day	Corpus callosum development
Strada	ENSMUSG00000069631.14.E17	-6.31	0.0000001	Tesa1-day	Epilepsy
E2f4	ENSMUSG00000014859.9.E10	-6.22	0.0000000	Tesa1-day	Cell cycle; cancer
Dopey2	ENSMUSG00000022946.10.E66	-6.19	0.0000000	Tesa1-day	Down syndrome
Gnb2	ENSMUSG00000029713.15.E71	-5.11	0.0000000	Tesa1-day	G protein

Table 5: Top 5 differentially used exons for each MSUS/TESA versus CTRL comparison that had significant results. Genes/exons were ordered by their $\log_2$ fold change and comparison. FDR was set to 0.1. Only unique genes are shown - if two or more exons from the same gene were in the top 5 results, the exon with the lowest adjusted P value was included in the table.

3.4.3 DEXSeq results for example genes

DEXSeq offers a useful plotting function that enables visualization of DEU results. In order to get a clear idea of which exons are being used differentially within a given gene, two exemplary genes, found in Table 5 and thus some of the top results, were used to generate DEXSeq plots Exons 123-128 from the gene Eif2ak4 (ENSMUSG00000005102) display DEU between MSUS

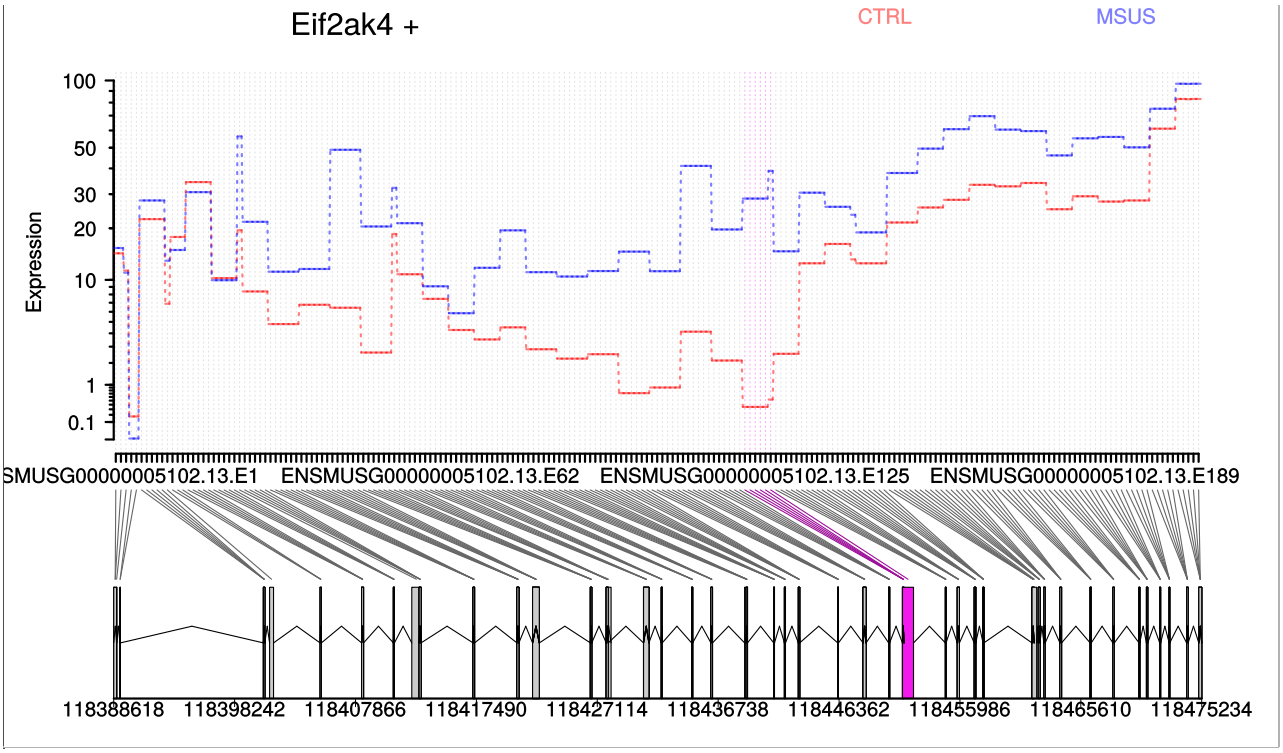

Figure 3.23: DEXSeq Plot - SC F1 Adult - Eif2ak4. Result plot of the expression of each exon of Eif2ak4 from the SC F1 Adult MSUS versus CTRL comparison. Differentially used exons between the two conditions are colored in pink.

and control in SC Adult samples. Figure 3.23 clearly shows that there is higher expression of these 6 exons in MSUS samples, which are labelled in pink. Although there appears to be a consistent trend of higher usage of most exons in MSUS samples, DGE analysis with limma/voom showed that there was no significant difference in gene expression between MSUS and control (fdrtool P value = 0.04). Exons 138-141, 176, 190-192, and 205-207 from the gene Fryl (ENSMUSG00000070733) display DEU between MSUS and control in Sperm samples. Figure 3.24 shows that in all significant exons there is lower expression of these exons in MSUS samples, which are labelled in pink. Unlike Eif2ak4, the overall expression of the exons in Fryl do not seem to be clearly higher in either MSUS or control samples. DGE analysis with limma/voom showed that there was no significant difference in gene expression between MSUS and control (fdrtool P value = 0.32).

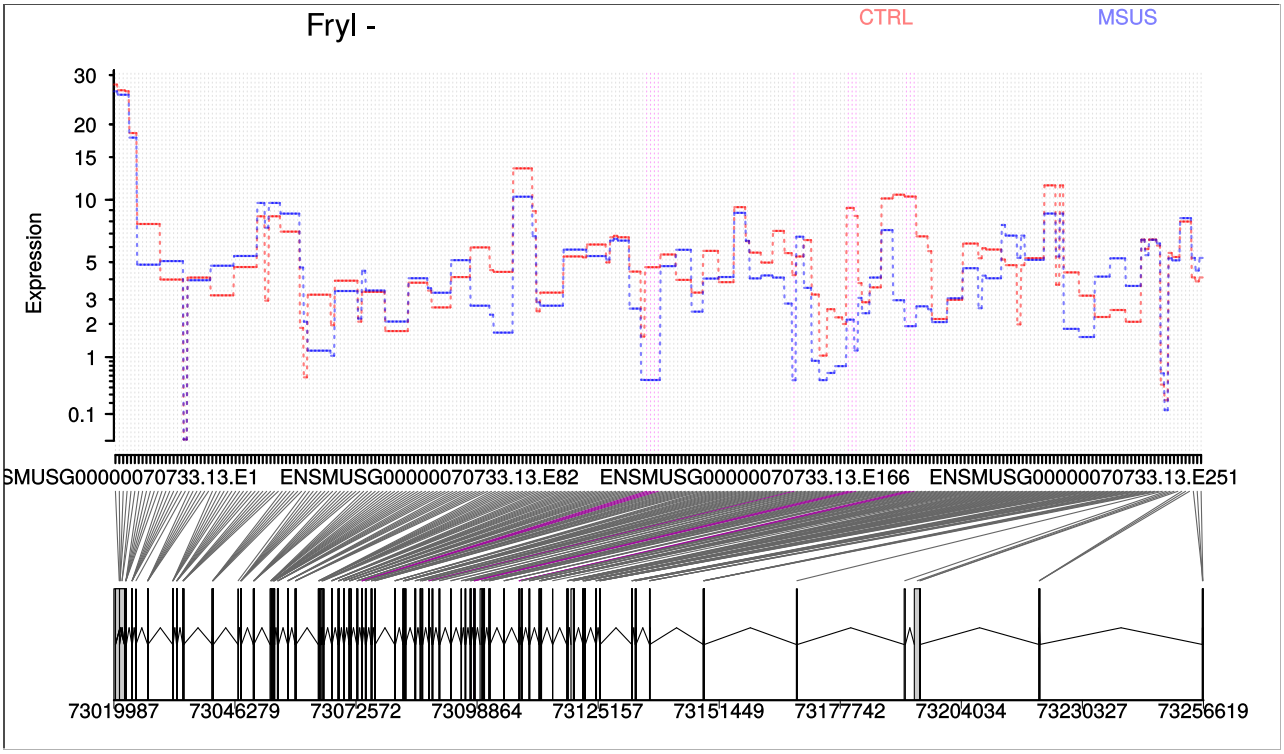

Figure 3.24: DEXSeq Plot - Sperm F1 - Fryl. Result plot of the expression of each exon of Fryl from the Sperm F1 MSUS versus CTRL comparison. Differentially used exons between the two conditions are colored in pink.

Chapter 4

Discussion

The MSUS paradigm has been demonstrated to induce a variety of changes in both the mice that underwent the MSUS treatment, as well as their offspring. In this study, computational analysis of RNA-Seq data from MSUS mice has revealed alterations in gene expression, pathway regulation, and exon usage. In all thirteen comparisons, a number of genes were found to be either up-regulated or down-regulated in MSUS mice versus control mice. Notably, there were high levels of DGE in both the SC Adult (228 DEG) and Sperm (193 DEG) comparisons. Within these significant genes, MSUS-associated genes were identified. Using the results of DGE, pathway analysis results showed extensive KEGG pathway enrichment in Liver F3 (128 - fGSEA) and Sperm (143 - fGSEA). Among the other comparisons, there were lower numbers of enriched pathways, and in some cases, no significant pathways were identified. Metabolic, neurological disease, and genetic information processing pathways were enriched in most cases. The results of the differential exon usage analysis were surprising - widespread differential exon usage was seen in SC PND8 (3,280 exons) and Tesa2day (13,090 exons).

All of these findings point to systematically, and in some cases large-scale, altered physiological functioning and regulation in germline and somatic tissues of MSUS mice. However, there is a need for more analysis to be done and for any findings to be experimentally validated in the laboratory by either *in vitro* or *in vivo* methods. For example, genes of interest such as the Pparg gene which is down-regulated in Zygote MSUS samples and Ctla2b which is upregulated in Liver F3 MSUS samples should have their expression levels validated in the laboratory. Both of these genes have been described in previous MSUS literature. In (Roszkowski et al., 2016), stress-induced expression of Ctla2b was repressed by injecting mice with propranolol, a chemical known to block the behavioral and molecular effects of stress, 30 minutes before exposure. The authors speculated that Ctla2b may act as a stress-induced immune cell activator since it mostly expressed in activated T-cells. Experiments run by Dr. Gretchen van Steenwyk, a postdoctoral researcher in the Mansuy laboratory, have shown that injection of mice with the PPAR-agonist tesaglitazar can mimic the increase of fatty acid metabolites seen in MSUS mice. The Pparg gene, which encodes the PPAR gamma receptor, is a direct target of tesaglitazar. Although the differential expression of this gene was seen only in the Zygote samples, the results from the tesaglitazar experiments demonstrate that there may be some connection between the metabolic effects of MSUS treatment and PPAR receptor activation.

Other genes found in the top DGE results are potentially connected to the known effects of the MSUS paradigm. A common characteristic of MSUS affected mice is an increase in body weight. Socs2, a gene that was found to be expressed significantly less in Liver F1 MSUS

samples, is associated with weight increase when expressed at lower than normal levels. Genes involved in learning and memory (Ntrk2 - Liver F1), hippocampal organization (Dclk2 - Liver F3), transcriptional regulation (Malat1 - Liver F3), epigenetic transcriptional activation (Pagr1a - Testis F1), DNA methylation repression (Pramel7 - SC PND8) and even histones (His2h2aa2 - Testis F1; H3f3c - Zygote) were found to be differentially expressed in MSUS mice. All of these genes present compelling targets for future research and validation.

It is of interest that the SC Adult comparison, which had over 200 DEGs, had no significantly enriched pathways in either CameraPR or fGSEA analysis. One possibility is that the genes that are DE happen to be distinctly less related to each other than the DEGs from other comparisons. Interestingly, the SC PND8 and PND15 comparisons, both of which had less DEGs, had enriched pathways.

The pathways that were shown to be altered/enriched in MSUS samples are intriguing and merit further investigation. Neurological disease pathways such as Alzheimer disease and Parkinson disease have differential regulation in some of the comparisons. Metabolic pathways like glutathione metabolism, steroid hormone biosynthesis, and amino acid biosynthesis all appear to be affected by MSUS treatment. The high numbers of altered pathways in the Liver F3 and Sperm comparisons is also noteworthy. With each having around 100 significantly enriched pathways, that comes to about one-third of the total KEGG pathways that were tested. This suggests that there are considerable changes in the functional profiles of these samples.

Another key observation is the widespread DEU in both the SC PND8 and Tesa1-day comparisons. In particular, the extensive DEU in SC PND8 samples seems to decrease by PND15 and in SC Adults, there are only a few dozen exons displaying DEU. In any case, there does not appear to be any clear patterns that link DEU with DGE. More importantly, the analysis of DEU in MSUS mice has not been done many times in the past, so it is still unclear how altered exon usage of genes, even MSUS-related genes, may impact the phenotype. Wet-lab experimental biology would be necessary to determine precisely the effects of these exonal changes. An intriguing possibility to explore is whether the high level of DEU can be attributed to altered methylation patterns. A recent study utilized CRISPR/Cas9 technology to demonstrate that methylation in exonal regions can lead to alternative splicing and differential exon inclusion (Shayevitch, Askayo, Keydar, & Ast, 2018).

4.1 Conclusion

The work presented in this thesis is one part of a larger effort to leverage the vast amounts of data from multi-omics experiments in an attempt to understand the mechanisms of epigenetic inheritance and TEI in mammals. The RNA-Seq data and its subsequent analysis presented here provides a strong foundation for informing future experiments and avenues of research. An important next step in the synthesis of a bioinformatic view of MSUS is to integrate different ‘omics experiments, including metabolomics, proteomics, and methylomics with the genomic/transcriptomic analysis completed here, in order to provide a more complete view of the effects the MSUS paradigm has on mice. The R package mixOmics, specifically the DIABLO framework of mixOmics, provides an excellent starting point to do this sort of multi-omic integrative analysis. By combining results from these various experiments together and performing a type of discriminant analysis, potential signatures and/or biomarkers of the MSUS

phenotype can be discovered.

Appendix A

Supplementary Data

A.1 Heatmap (combined)

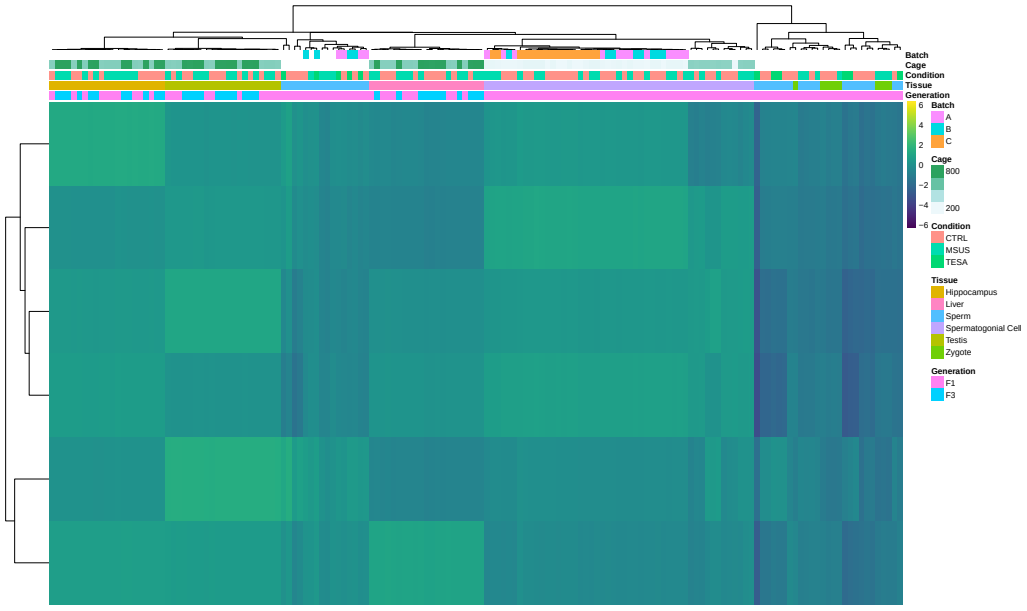

Figure A.1: Heatmap - All Samples (intersect). Heatmap of log2 Counts-per-million of gene counts from all 155 samples across the 7 datasets. The samples were combined based on genes that were reported in the limma results of all 7 datasets. A total of 3301 genes were common between the datasets. Rows and columns were clustered. Genes were clustered using k-means clustering, with k = 6. Samples are annotated with information including batch, cage, condition, tissue, and generation.

A.2 KEGG Pathway Dot Plots

The following plots display some of the top-enriched pathways for each comparison. The x-axis range is different in each plot.

A.2.1 Somatic + Testis

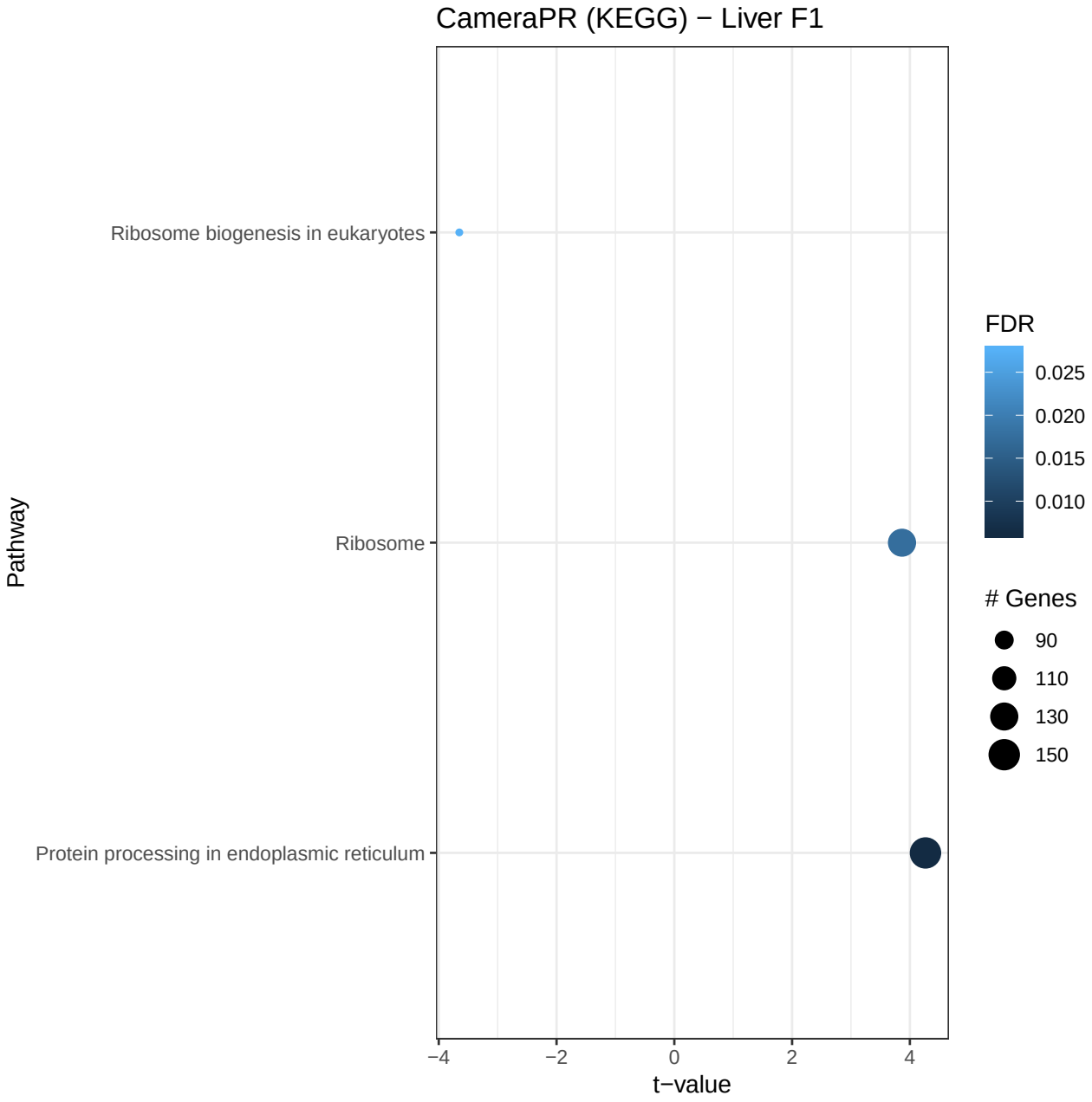

Figure A.2: Dot plot of Liver F1 CameraPR pathway analysis results. The t-values of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

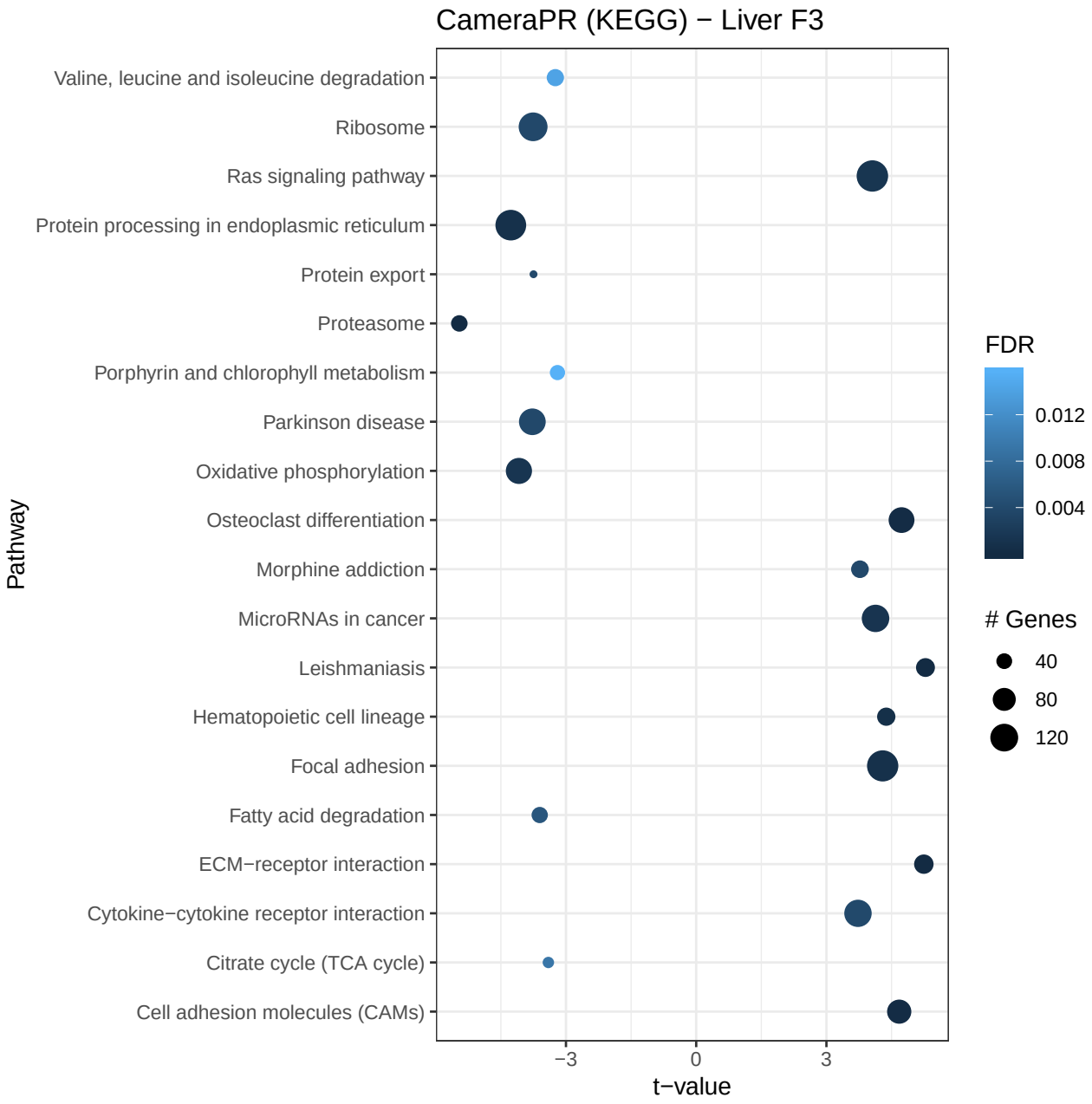

Figure A.3: Dot plot of Liver F3 CameraPR pathway analysis results. The t-values of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

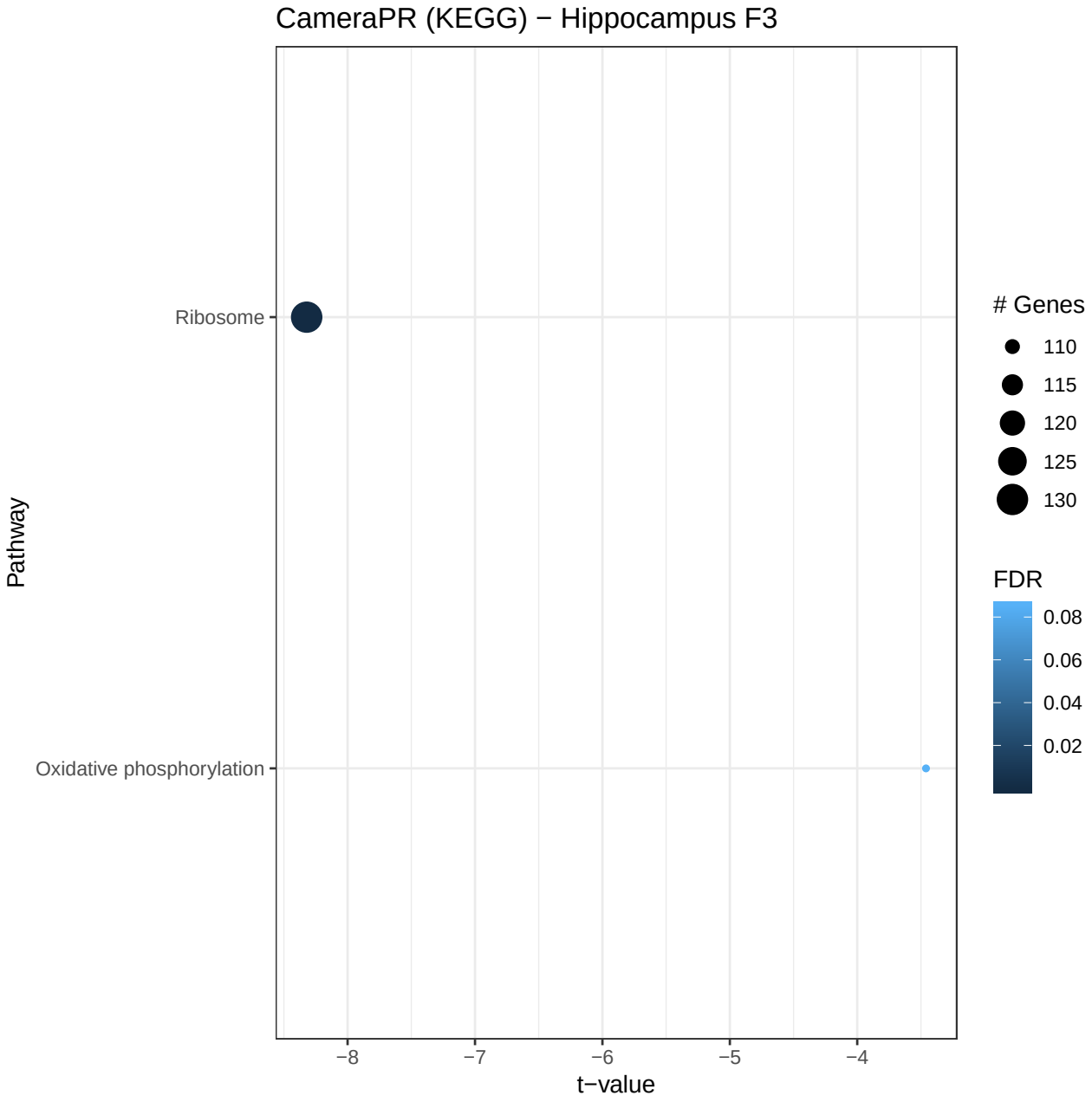

Figure A.4: Dot plot of Hippocampus F3 CameraPR pathway analysis results. The t-values of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

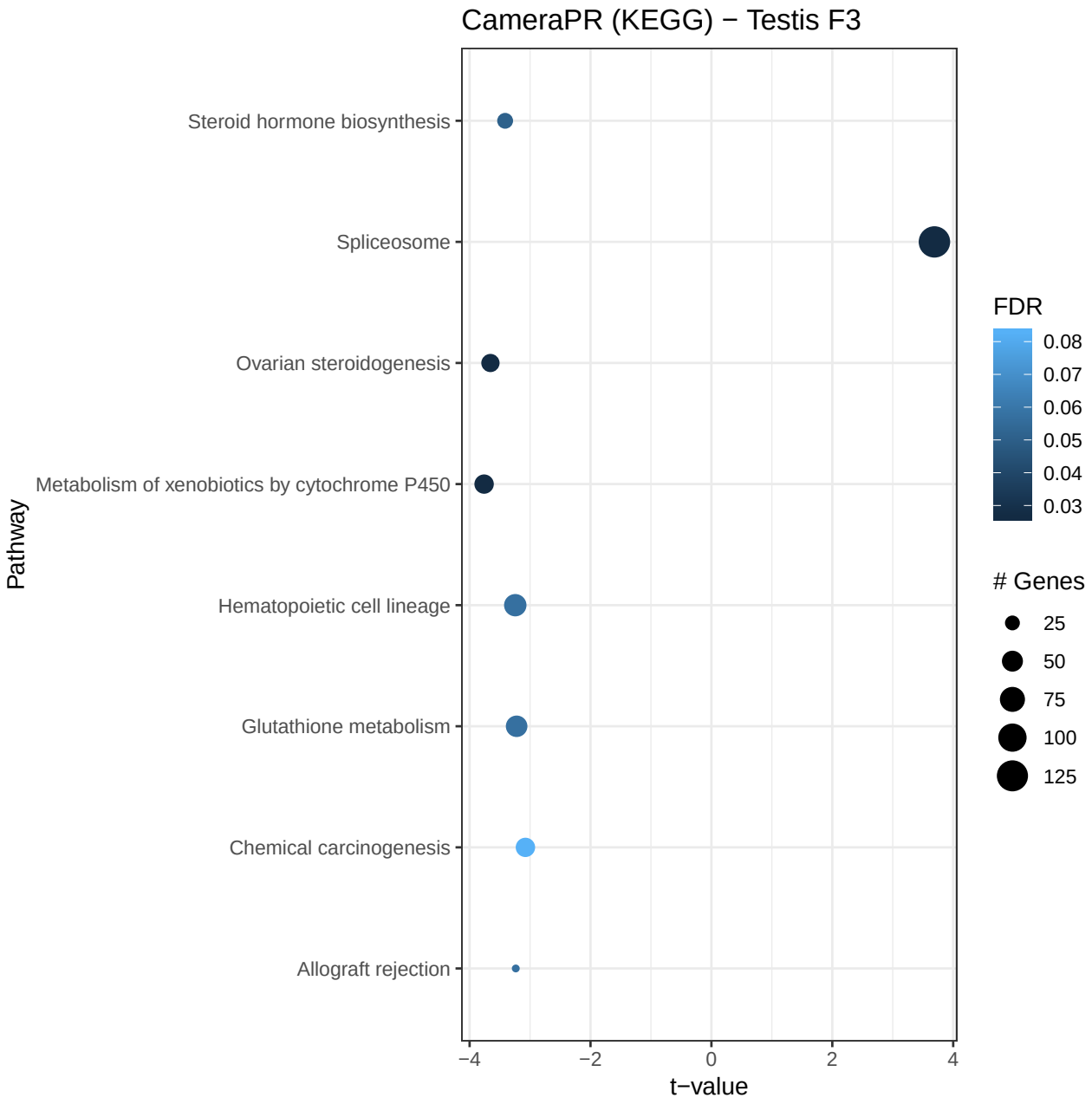

Figure A.5: Dot plot of Testis F3 CameraPR pathway analysis results. The t-values of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

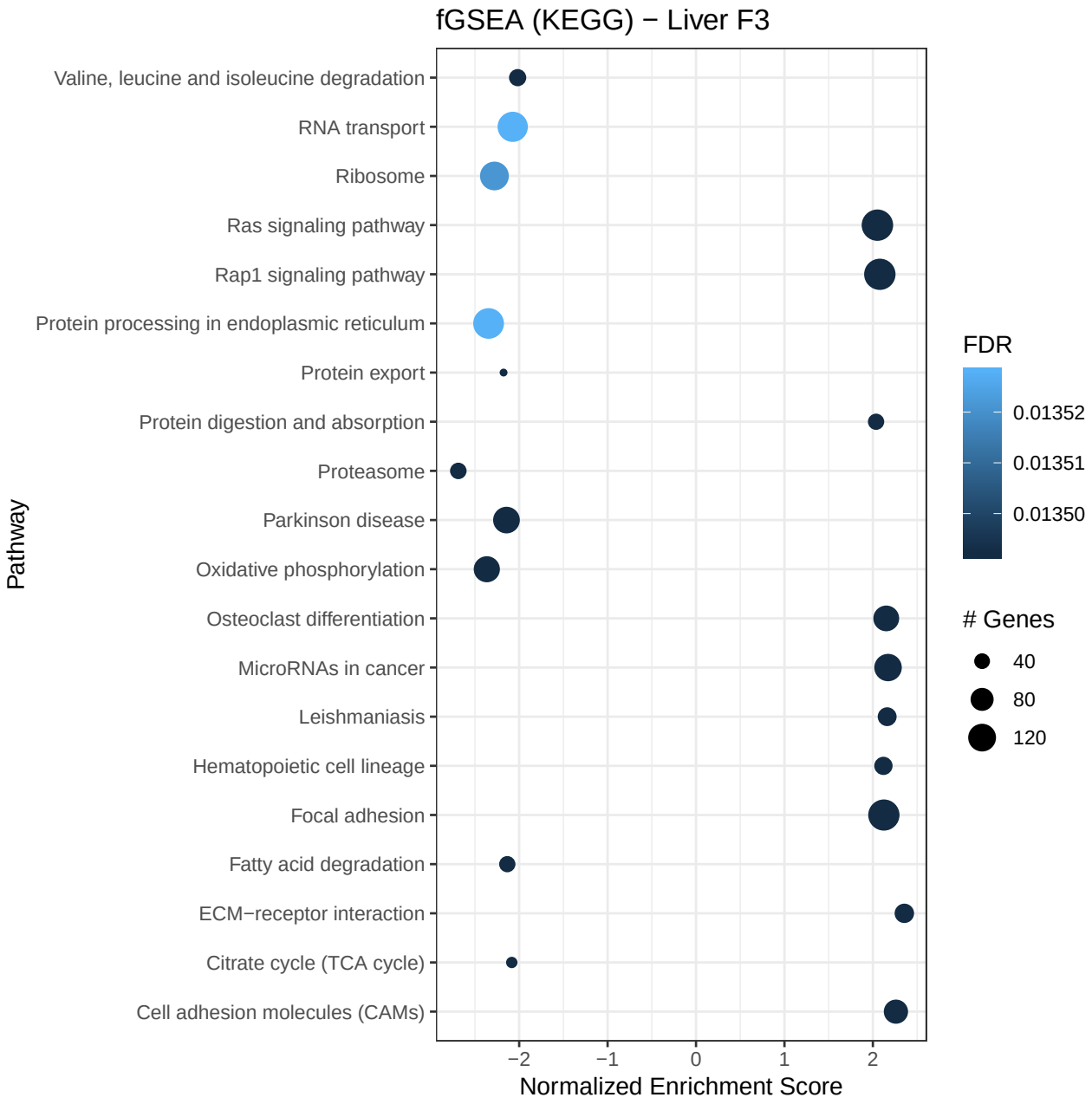

Figure A.6: Dot plot of Liver F3 fGSEA pathway analysis results. The normalized enrichment score of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

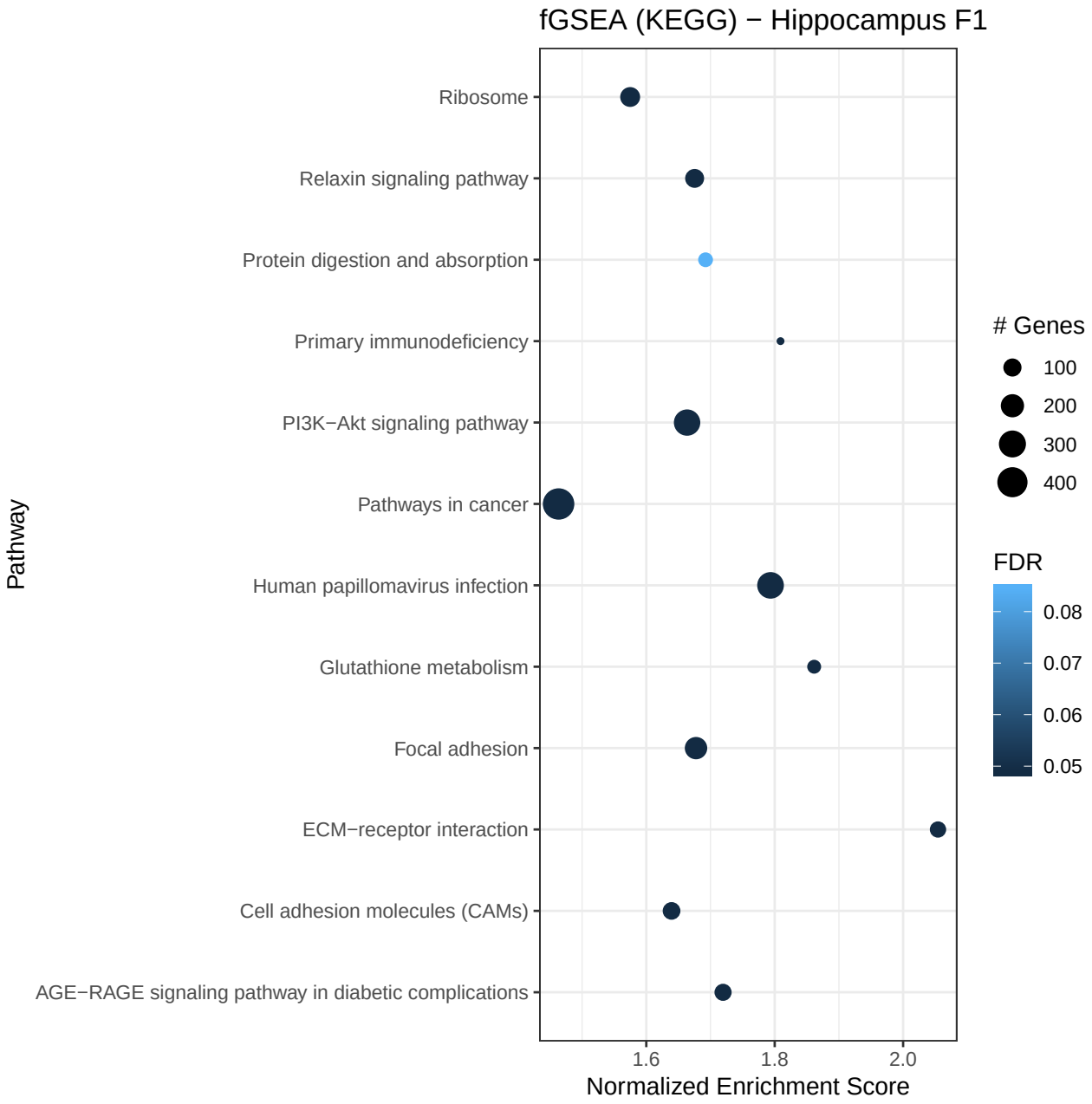

Figure A.7: Dot plot of Hippocampus F1 fGSEA pathway analysis results. The normalized enrichment score of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

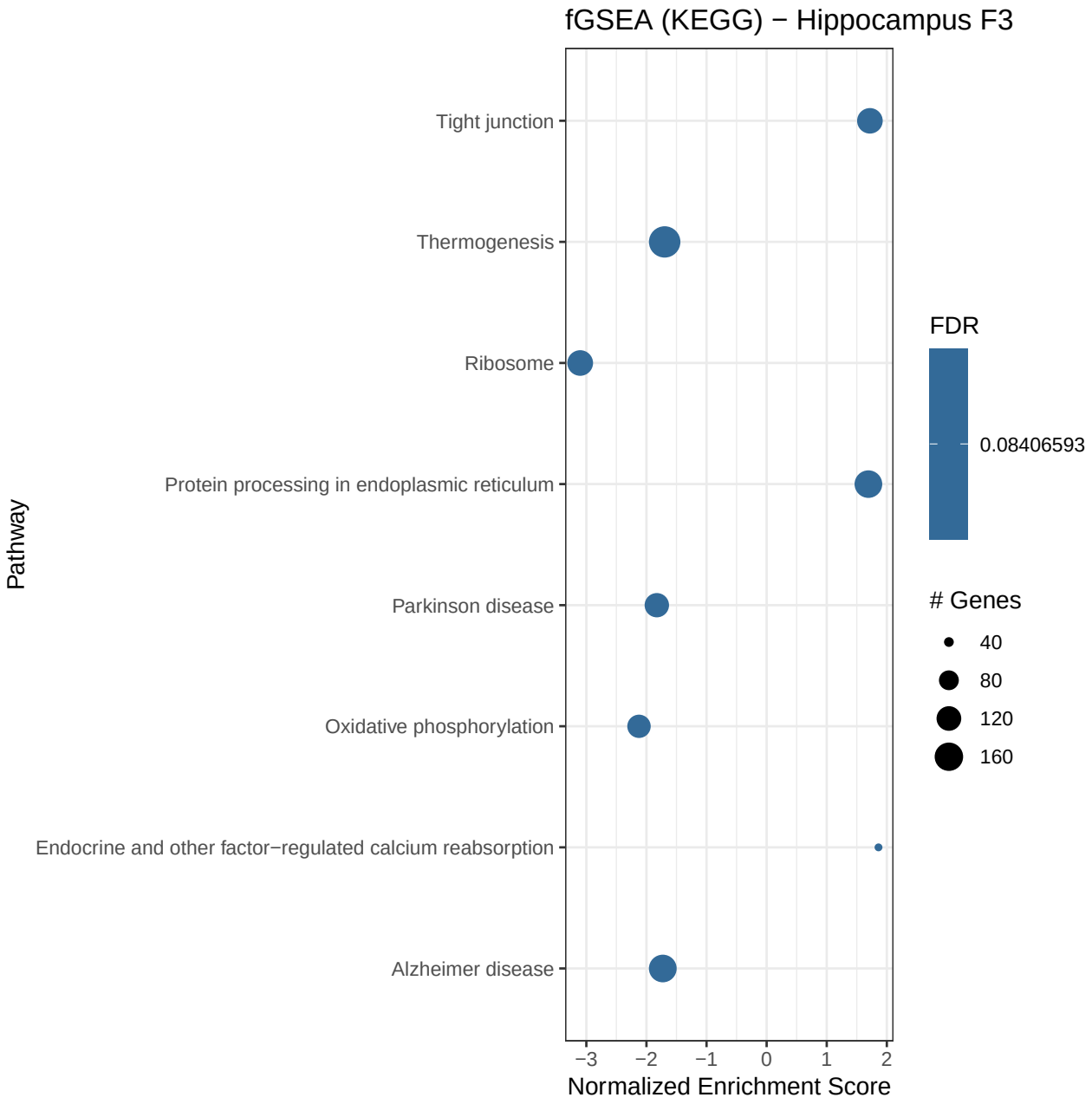

Figure A.8: Dot plot of Hippocampus F3 fGSEA pathway analysis results. The normalized enrichment score of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

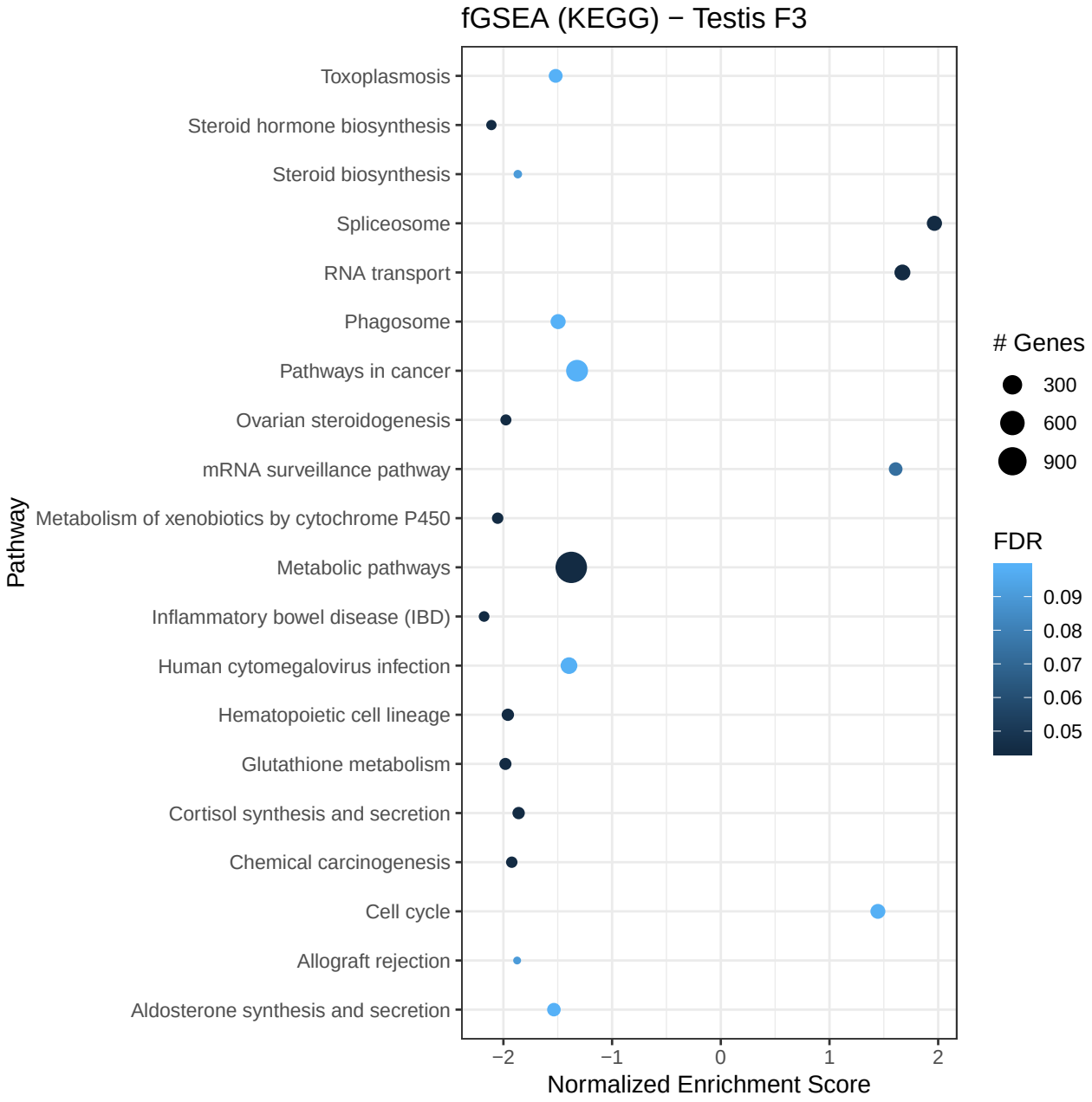

Figure A.9: Dot plot of Testis F3 fGSEA pathway analysis results. The normalized enrichment score of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

A.2.2 SC PND8

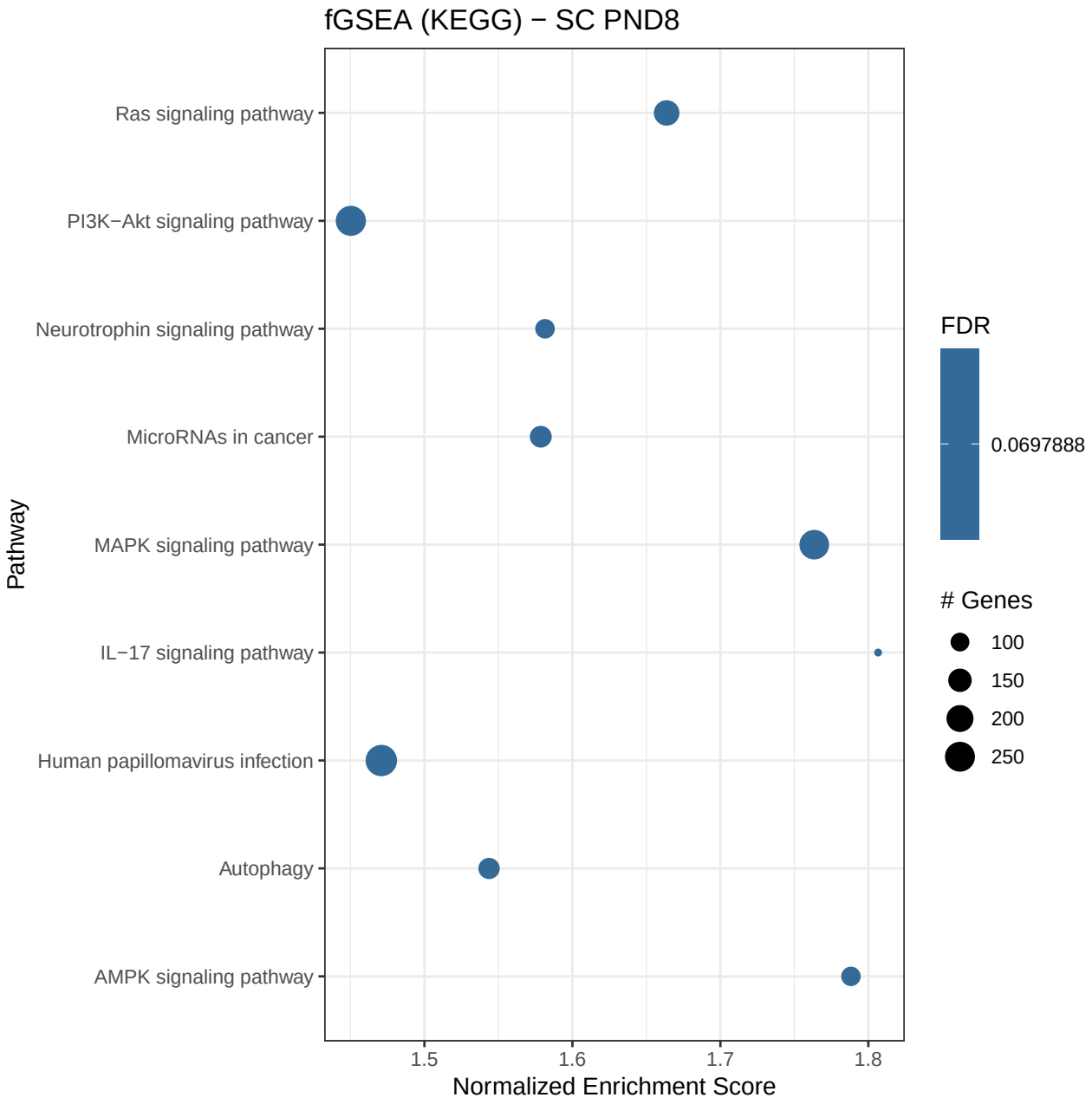

Figure A.10: Dot plot of SC PND8 fGSEA pathway analysis results. The normalized enrichment score of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

A.2.3 SC PND15

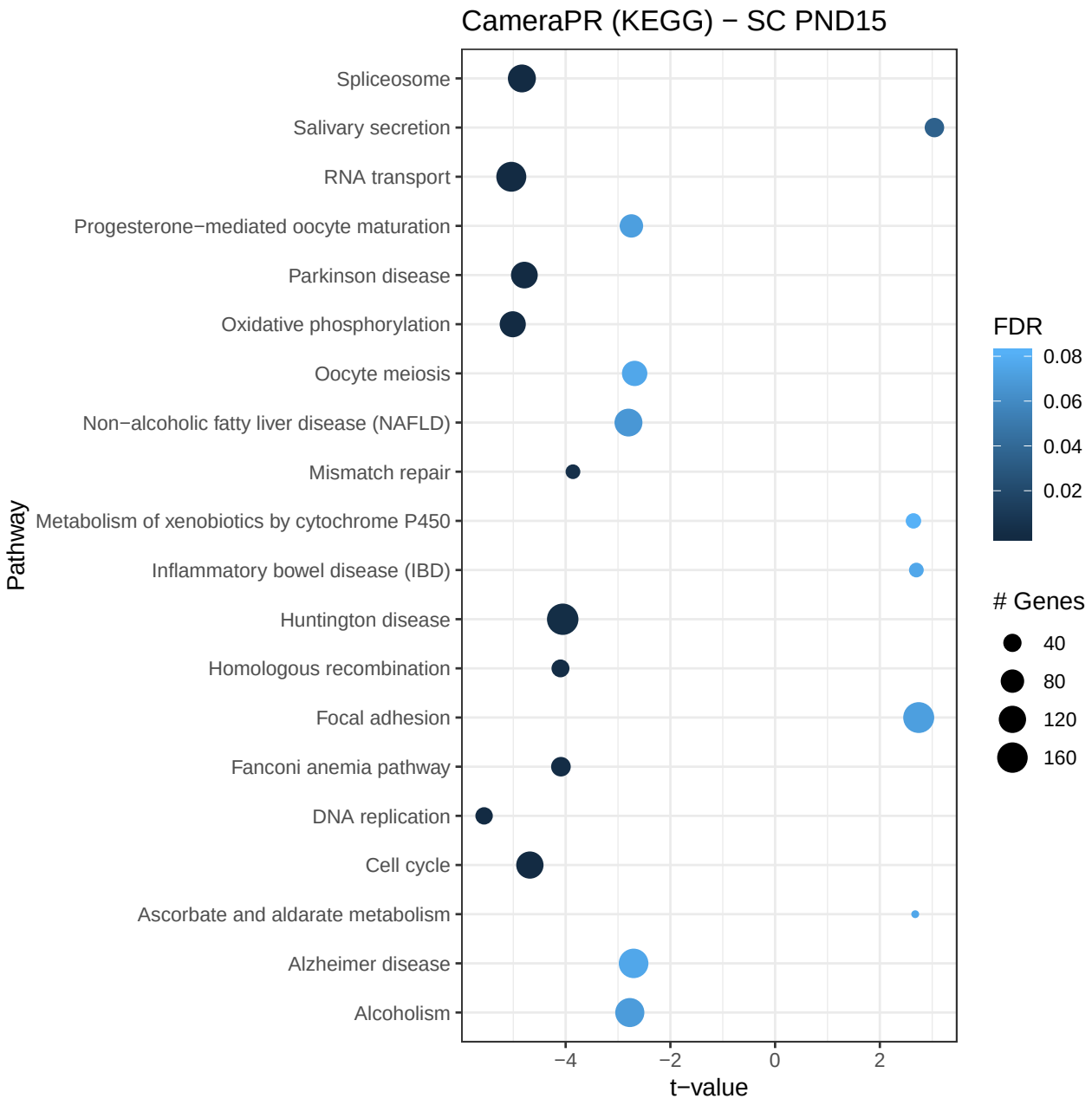

Figure A.11: Dot plot of SC PND15 fGSEA pathway analysis results. The t-values of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

Figure A.12: Dot plot of SC PND15 fGSEA pathway analysis results. The normalized enrichment score of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

A.2.4 Sperm

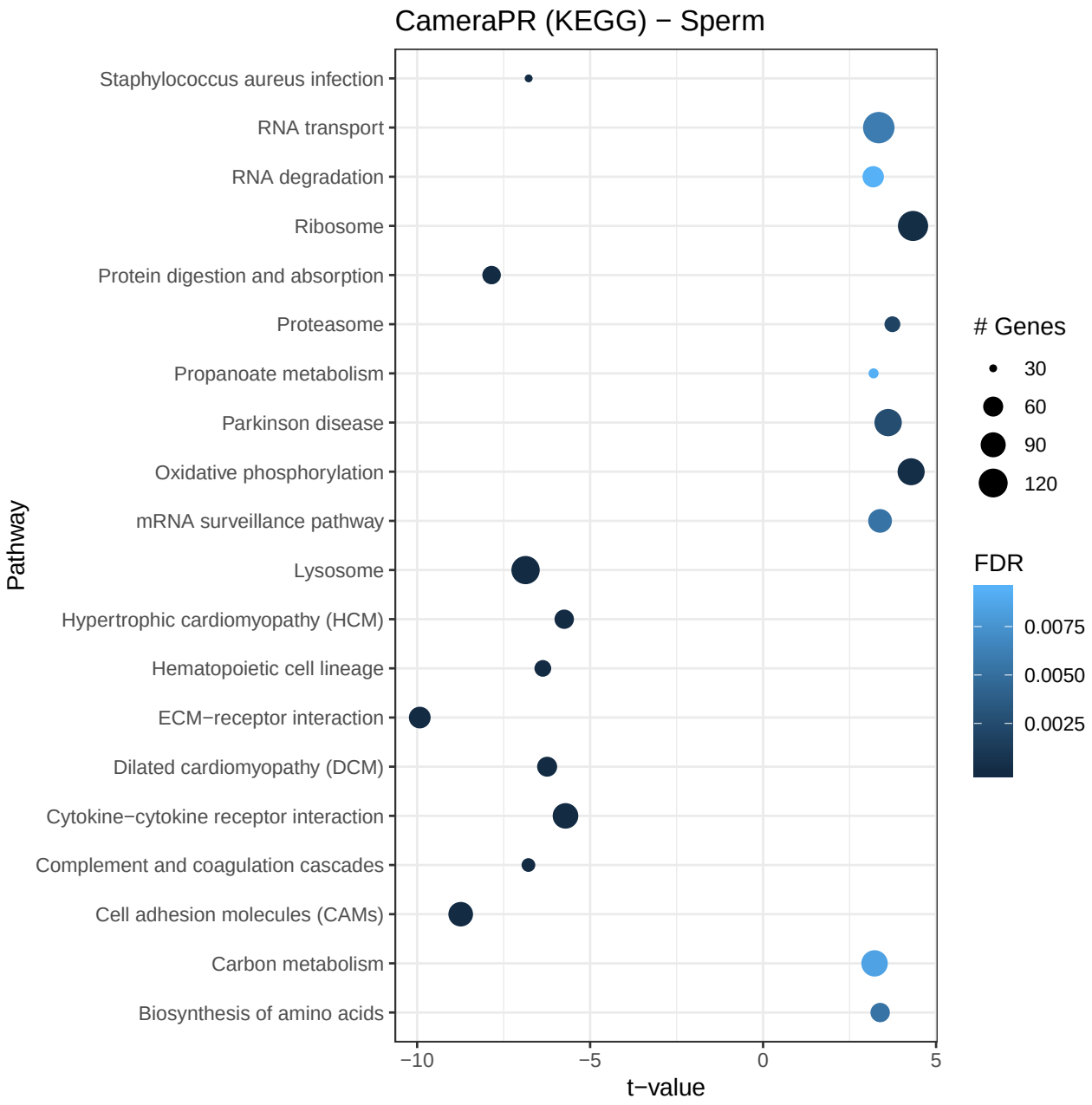

Figure A.13: Dot plot of Sperm fGSEA pathway analysis results. The t-values of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

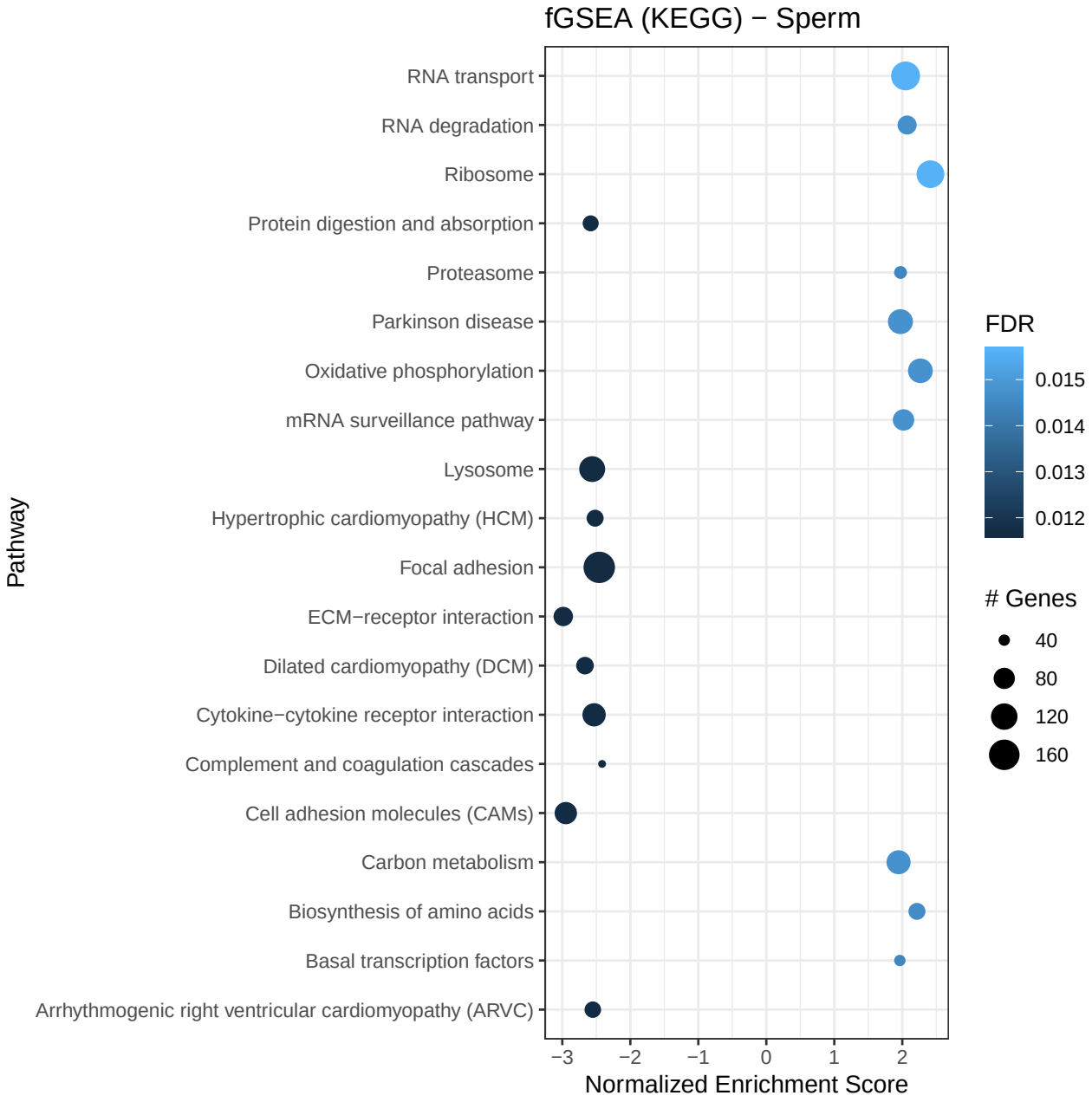

Figure A.14: Dot plot of Sperm fGSEA pathway analysis results. The normalized enrichment score of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

A.2.5 Zygote

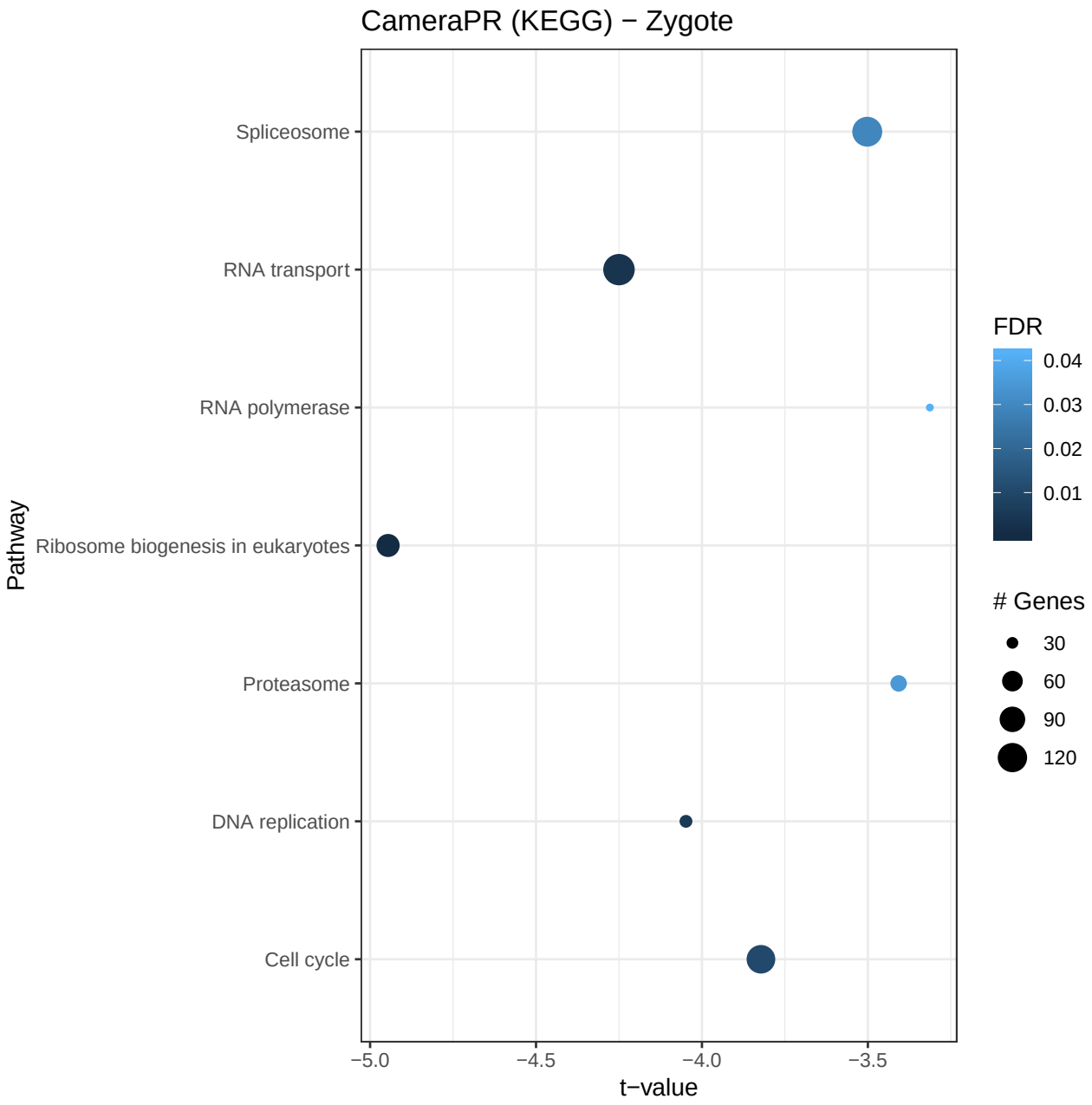

Figure A.15: Dot plot of Zygote fGSEA pathway analysis results. The t-values of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

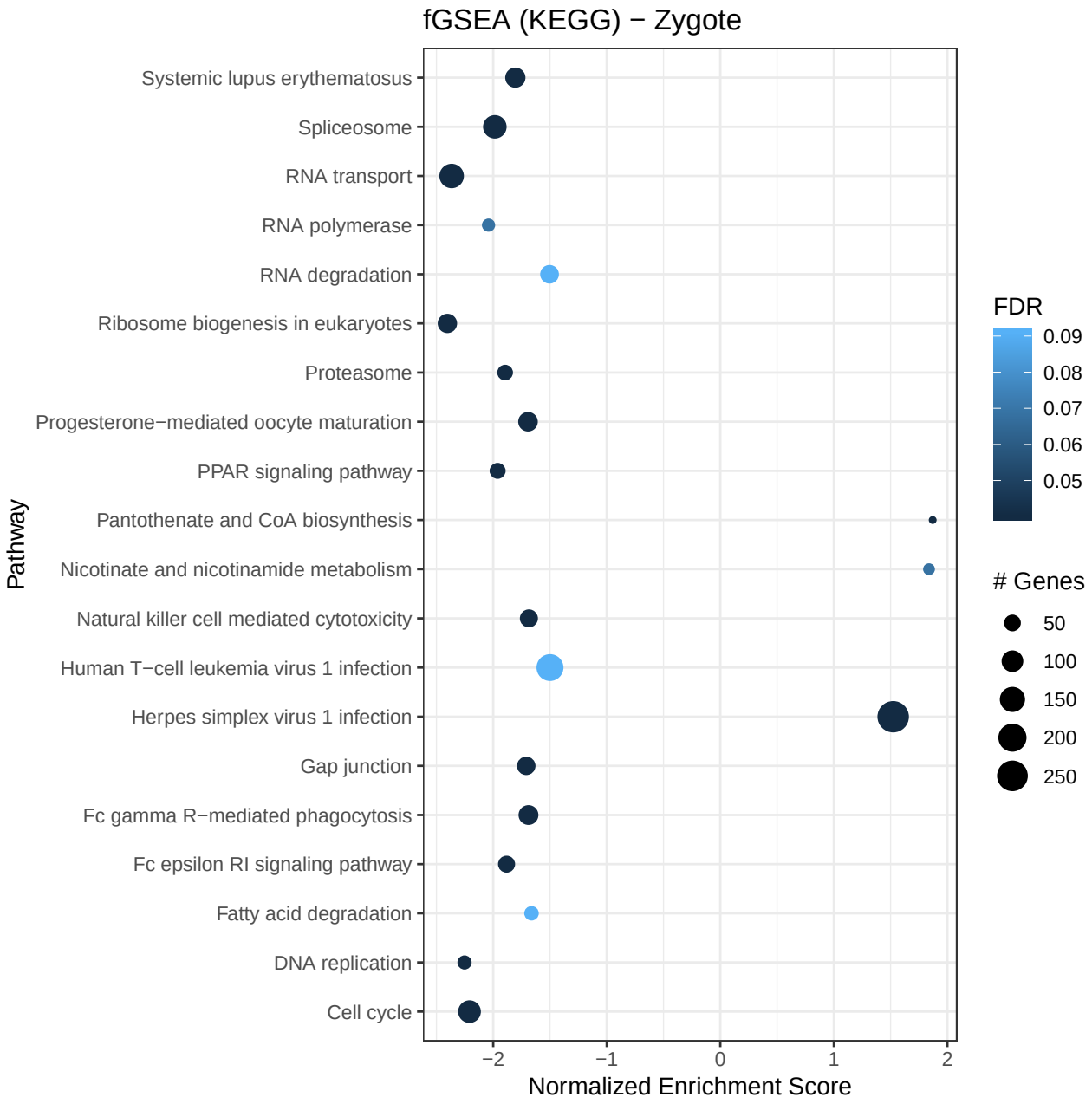

Figure A.16: Dot plot of Zygote fGSEA pathway analysis results. The normalized enrichment score of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

A.2.6 Tesa46-day

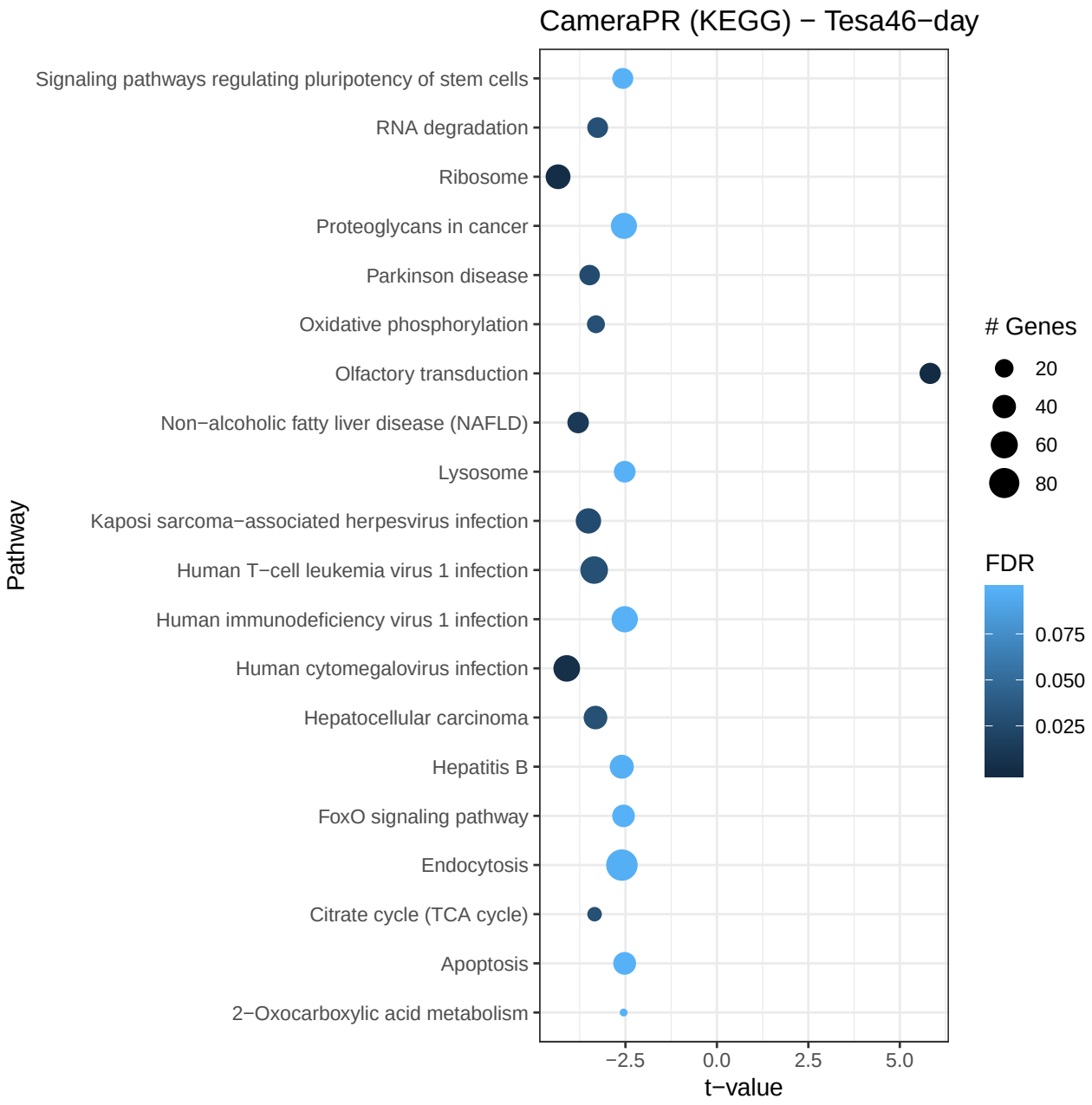

Figure A.17: Dot plot of Tesa46-day fGSEA pathway analysis results. The t-values of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

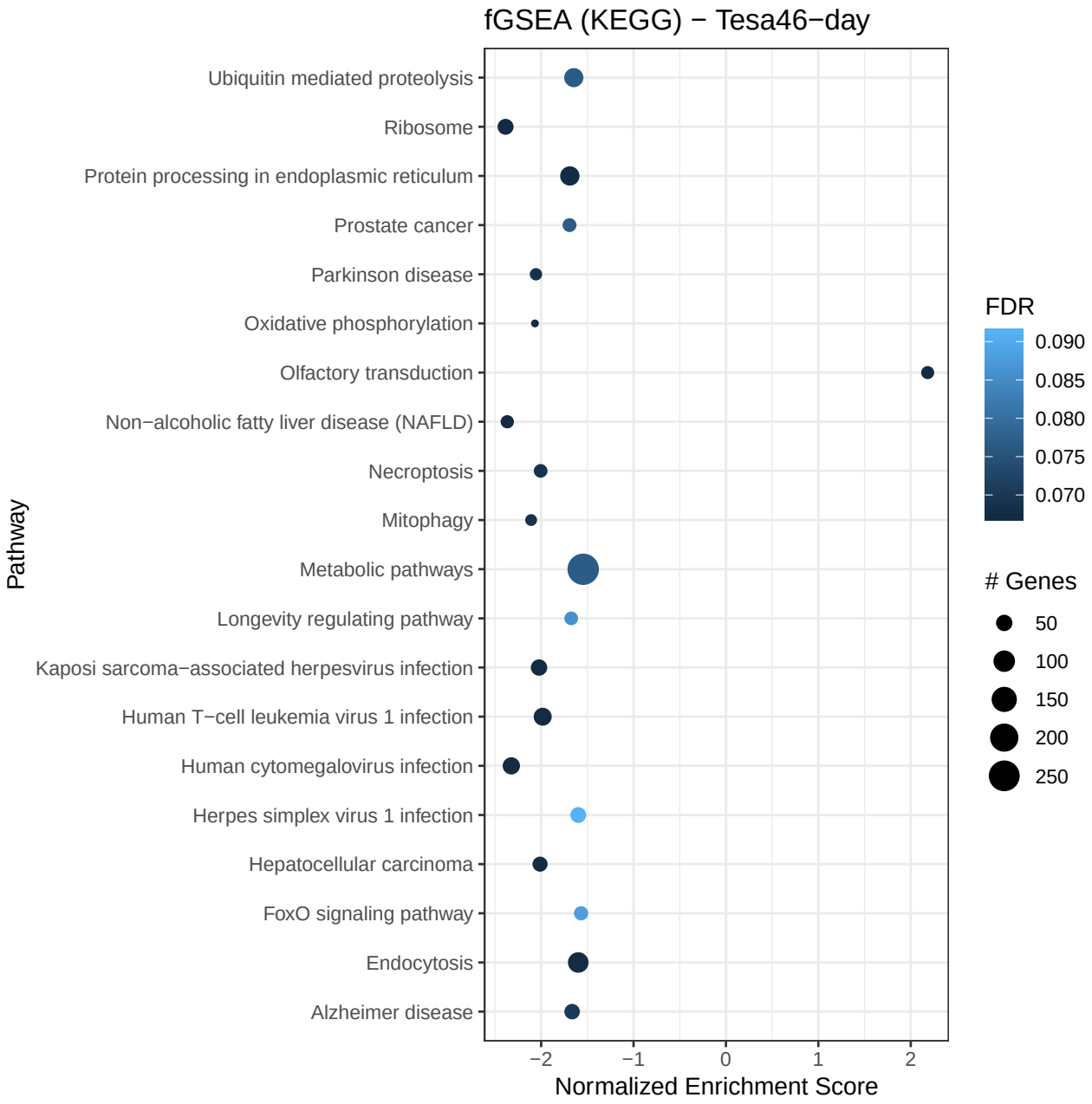

Figure A.18: Dot plot of Tesa46-day fGSEA pathway analysis results. The normalized enrichment score of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

A.2.7 Tesa1-day

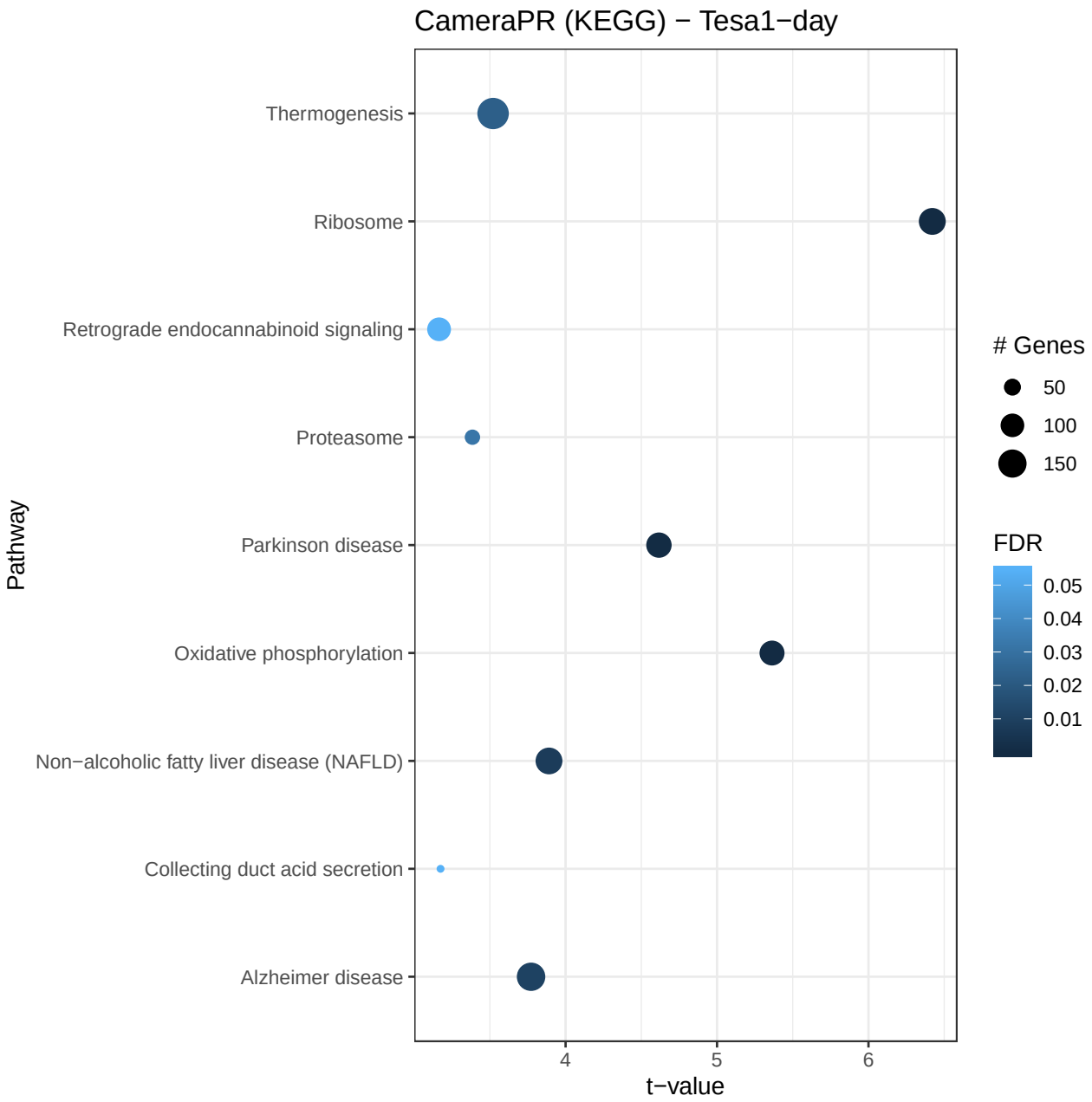

Figure A.19: Dot plot of Tesa1-day fGSEA pathway analysis results. The t-values of the top up-regulated and down-regulated pathways with an FDR of less than 0.1 are plotted. The size of the dot corresponds to the size of the pathway gene set. The color of the dot denotes the FDR value. Pathways are in reverse alphabetical order.

References

Allis, C. D., & Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. *Nature Reviews Genetics*, *17*(8), 487–500. http://doi.org/10.1038/nrg.2016.59

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. http://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. http://doi.org/10.1101/gr.133744.111

Babb, J. A., Carini, L. M., Spears, S. L., & Nephew, B. C. (2014). Transgenerational effects of social stress on social behavior, corticosterone, oxytocin, and prolactin in rats. *Hormones and Behavior*, *65*(4), 386–393. http://doi.org/10.1016/j.yhbeh.2014.03.005

Benevento, M., Iacono, G., Selten, M., Ba, W., Oudakker, A., Frega, M., ... Kasri, N. N. (2016). Histone methylation by the kleefstra syndrome protein EHMT1 mediates homeostatic synaptic scaling. *Neuron*, *91*(2), 341–355. http://doi.org/10.1016/j.neuron.2016.06.003

Berge, K. V. D., Hembach, K., Soneson, C., Tiberi, S., Clement, L., Love, M. I., ... Robinson, M. (2018). RNA sequencing data: Hitchhikers guide to expression analysis. http://doi.org/10.7287/peerj.preprints.27283v2

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & Development*, *16*(1), 6–21. http://doi.org/10.1101/gad.947102

Bohacek, J., & Mansuy, I. M. (2015). Molecular insights into transgenerational non-genetic inheritance of acquired behaviours. *Nature Reviews Genetics*, *16*(11), 641–652. http://doi.org/10.1038/nrg3964

Bumgarner, R. (2013). Overview of dna microarrays: Types, applications, and their future. *Current Protocols in Molecular Biology*. http://doi.org/10.1002/0471142727.mb2201s101

Can, A., Dao, D. T., Arad, M., Terrillion, C. E., Piantadosi, S. C., & Gould, T. D. (2011). The mouse forced swim test. *Journal of Visualized Experiments*, (58). http://doi.org/10.3791/3638

Castillo-Davis, C. I., & Hartl, D. L. (2003). GeneMerge–post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, *19*(7), 891–892. http://doi.org/10.1093/bioinformatics/btg114

Dick, A., & Provencal, N. (2018). Central neuroepigenetic regulation of the hypothalamicPituitaryAdrenal axis. In *Progress in molecular biology and translational science* (pp. 105–127).

Elsevier. http://doi.org/10.1016/bs.pmbts.2018.04.006

Franklin, T. B., Russig, H., Weiss, I. C., Gräff, J., Linder, N., Michalon, A., ... Mansuy, I. M. (2010). Epigenetic transmission of the impact of early stress across generations. *Biological Psychiatry*, *68*(5), 408–415. http://doi.org/10.1016/j.biopsych.2010.05.036

Froussios, K., Mourão, K., Simpson, G. G., Barton, G. J., & Schurch, N. J. (2017). Identifying differential isoform abundance with RATs: A universal tool and a warning. http://doi.org/10.1101/132761

Gapp, K., Jawaid, A., Sarkies, P., Bohacek, J., Pelczar, P., Prados, J., ... Mansuy, I. M. (2014a). Implication of sperm RNAs in transgenerational inheritance of the effects of early trauma in mice. *Nature Neuroscience*, *17*(5), 667–669. http://doi.org/10.1038/nn.3695

Gapp, K., Soldado-Magraner, S., Alvarez-Sánchez, M., Bohacek, J., Vernaz, G., Shu, H., ... Mansuy, I. M. (2014b). Early life stress in fathers improves behavioural flexibility in their offspring. *Nature Communications*, *5*(1). http://doi.org/10.1038/ncomms6466

Geistlinger, L., Csaba, G., Kuffner, R., Mulder, N., & Zimmer, R. (2011). From sets to graphs: Towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics*, *27*(13), i366–i373. http://doi.org/10.1093/bioinformatics/btr228

Hanauer, A. (2002). Coffin-lowry syndrome: Clinical and molecular features. *Journal of Medical Genetics*, *39*(10), 705–713. http://doi.org/10.1136/jmg.39.10.705

Huntzinger, E., & Izaurralde, E. (2011). Gene silencing by microRNAs: Contributions of translational repression and mRNA decay. *Nature Reviews Genetics*, *12*(2), 99–110. http://doi.org/10.1038/nrg2936

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, *8*(2), e1002375. http://doi.org/10.1371/journal.pcbi.1002375

Lacal, I., & Ventura, R. (2018). Epigenetic inheritance: Concepts, mechanisms and perspectives. *Frontiers in Molecular Neuroscience*, *11*. http://doi.org/10.3389/fnmol.2018.00292

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29. http://doi.org/10.1186/gb-2014-15-2-r29

Li, E., Bestor, T. H., & Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, *69*(6), 915–926. http://doi.org/10.1016/0092-8674(92)90611-f

Liao, Y., Smyth, G. K., & Shi, W. (2019). The r package rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, *47*(8), e47–e47. http://doi.org/10.1093/nar/gkz114

Lind, M. I., & Spagopoulou, F. (2018). Evolutionary consequences of epigenetic inheritance. *Heredity*, *121*(3), 205–209. http://doi.org/10.1038/s41437-018-0113-y

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). http://doi.org/10.1186/s13059-014-0550-8

Mannironi, C., Camon, J., Vito, F. D., Biundo, A., Stefano, M. E. D., Persiconi, I., … Presutti, C. (2013). Acute stress alters amygdala microRNA miR-135a and miR-124 expression: Inferences for corticosteroid dependent stress response. *PLoS ONE*, *8*(9), e73385. http://doi.org/10.1371/journal.pone.0073385

Matosin, N., Cruceanu, C., & Binder, E. B. (2017). Preclinical and clinical evidence of DNA methylation changes in response to trauma and chronic stress. *Chronic Stress*, *1*, 247054701771076. http://doi.org/10.1177/2470547017710764

Nativio, R., Donahue, G., Berson, A., Lan, Y., Amlie-Wolf, A., Tuzer, F., … Berger, S. L. (2018). Dysregulation of the epigenetic landscape of normal aging in alzheimer's disease. *Nature Neuroscience*, *21*(4), 497–505. http://doi.org/10.1038/s41593-018-0101-9

Panning, B., & Jaenisch, R. (1996). DNA hypomethylation can activate xist expression and silence x-linked genes. *Genes & Development*, *10*(16), 1991–2002. http://doi.org/10.1101/gad.10.16.1991

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. http://doi.org/10.1038/nmeth.4197

Provençal, N., & Binder, E. B. (2015). The effects of early life stress on the epigenome: From the womb to adulthood and even before. *Experimental Neurology*, *268*, 10–20. http://doi.org/10.1016/j.expneurol.2014.09.001

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

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. http://doi.org/10.1093/bioinformatics/btp616

Roszkowski, M., Manuella, F., Ziegler, L. von, Durán-Pacheco, G., Moreau, J.-L., Mansuy, I. M., & Bohacek, J. (2016). Rapid stress-induced transcriptomic changes in the brain depend on beta-adrenergic signaling. *Neuropharmacology*, *107*, 329–338. http://doi.org/10.1016/j.neuropharm.2016.03.046

Salomonis, N., Hanspers, K., Zambon, A. C., Vranizan, K., Lawlor, S. C., Dahlquist, K. D., … Pico, A. R. (2007). GenMAPP 2: New features and resources for pathway analysis. *BMC Bioinformatics*, *8*(1), 217. http://doi.org/10.1186/1471-2105-8-217

Sergushichev, A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. http://doi.org/10.1101/060012

Shayevitch, R., Askayo, D., Keydar, I., & Ast, G. (2018). The importance of DNA methylation of exons on alternative splicing. *RNA*, *24*(10), 1351–1362. http://doi.org/10.1261/rna.064865.117

Skvortsova, K., Iovino, N., & Bogdanović, O. (2018). Functions and mechanisms of epigenetic inheritance in animals. *Nature Reviews Molecular Cell Biology*, *19*(12), 774–790. http://doi.org/10.1038/s41580-018-0074-2

Steenwyk, G. van, Roszkowski, M., Manuella, F., Franklin, T. B., & Mansuy, I. M. (2018). Transgenerational inheritance of behavioral and metabolic effects of paternal exposure to traumatic stress in early postnatal life: Evidence in the 4th generation. *Environmental Epigenetics*, *4*(2). http://doi.org/10.1093/eep/dvy023

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. http://doi.org/10.1073/pnas.0506580102

Suzuki, M. M., & Bird, A. (2008). DNA methylation landscapes: Provocative insights from epigenomics. *Nature Reviews Genetics*, *9*(6), 465–476. http://doi.org/10.1038/nrg2341

Sweatt, J. D. (2013). The emerging field of neuroepigenetics. *Neuron*, *80*(3), 624–632. http://doi.org/10.1016/j.neuron.2013.10.023

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., … Romero, R. (2008). A novel signaling pathway impact analysis. *Bioinformatics*, *25*(1), 75–82. http://doi.org/10.1093/bioinformatics/btn577

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2012). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, *31*(1), 46–53. http://doi.org/10.1038/nbt.2450

Waddington, C. H. (1942). The epigenotype. *International Journal of Epidemiology*, *41*(1), 10–13. http://doi.org/10.1093/ije/dyr184

Walf, A. A., & Frye, C. A. (2007). The use of the elevated plus maze as an assay of anxiety-related behavior in rodents. *Nature Protocols*, *2*(2), 322–328. http://doi.org/10.1038/nprot.2007.44

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. http://doi.org/10.1038/nrg2484

Wu, D., & Smyth, G. K. (2012). Camera: A competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, *40*(17), e133–e133. http://doi.org/10.1093/nar/gks461

Zovkic, I. B., Guzman-Karlsson, M. C., & Sweatt, J. D. (2013). Epigenetic regulation of memory formation and maintenance. *Learning & Memory*, *20*(2), 61–74. http://doi.org/10.1101/lm.026575.112