**PROJECT 3 Final Report**

Nkob; Heather Baldwin, Ida Holásková, Denzel Middleton

December 20, 2019

**High-throughput sequencing of mRNA from murine ovaries exposed to propanil**

**INTRODUCTION**

  Propanil (3,4-dichloropropionanilide, DCPA) is a selective herbicide used in rice paddies in the United States and around the world (*1*). In susceptible plants, for example, barnyard grass, propanil inhibits photosynthesis and induces cell death (*2*). The same enzymatic machinery that rice plants utilize to metabolize propanil to DCA and propionic acid is present in mammals (*3*). However, in mammals, both the parent compound and some further metabolites have undesirable physiological consequences. Specifically, after an acute high dose exposure, the central nervous, circulatory, immune as well as reproductive systems, including pregnancy losses are affected (*3*–*6*). An interesting finding in mice exposed to propanil was that it led to a many-fold increase of antigen-specific antibody-producing cells in the spleen. Moreover, this effect depended on intact ovaries, as their removal revoked the immune response (*7*). The current working model for studying how the immune response is modified by propanil includes the injection of an antigen, *Streptococcus pneumoniae* (HKSP) along with propanil (*7*, *8*). To investigate the unexpected connection of spleen and ovaries, an important step is to determine the effects of propanil exposure on ovaries. To meet this end, the current study utilized female mice which were challenged with HKSP and concurrently exposed to propanil or peanut oil (control). Ovaries were then collected 24 hours post-exposure. The aim of this project was to utilize RNA Sequencing (RNA Seq) data to determine differences in global gene expression of mice treated with propanil and HKSP as well as mice only challenged with HKSP (control).

**RESPONSE TO REVIEWERS**

Authors appreciate the time and efforts put forth by reviewers. The innovative component of this experiment was in fact the technology being used to answer an innovative application of the effects of 24-hour propanil exposure to the ovarian transcriptome. The division of labor sought out heavy brainstorming, all group members wanted to reach a level of proficiency in analyzing RNA-sequencing data. To ensure this, we divided the samples and gave hard deadlines for pipeline steps to be completed by. Each member was given four samples to complete up to the Ballgown step in the pipeline. Once at *Stringtie*, we created the merged file and copied it (as well as other necessary sample files) to each of our scratch folders. This allowed all group members to run through the pipeline in its entirety. We appreciate the suggestions for expected outcomes and inferences on data analysis and reporting.

**RESULTS**

*Data Retrieval*

To test the experimental hypothesis that ovarian gene expression is modified after propanil exposure, paired end Illumina sequencing data was utilized from a graduate student in the Department of Animal & Nutritional Sciences at West Virginia University. The study included ovarian RNA isolation from 12 HKSP-challenged female mice (*Mus musculus,* strain C57Bl/6J, The Jackson Laboratory, Bar Harbor, ME) as six mice were treated with propanil and six were treated with peanut oil (control). Raw data in the .fastq file format were downloaded from the sequencing core web page. Raw reads were mapped to the reference genome (Accession: GCA\_000001635.8), which is available from NCBI’s assembly platform in the .fna, .gft and .gff file formats. This reference genome assembly has 2,730,855,475 bp with 605 contigs and 162 scaffolds.

*Fastqc (v0.11.7)*

Fastqc was utilized to individually check the quality of all samples (12 paired end samples= 24 files). Overall, the quality of reads was high; all 24 samples had the average quality score equal to or above 32. There were no N content and zero adapter content for all 24 samples. Genome, as well as transcriptome coverage were calculated (**Table 1**), using the *Fastqc* resultswith respect to the size of reference genome and size of the mouse transcriptome, respectively. Transcriptome coverage was sufficiently large. The only warning we received was a finding of 1-3 specific types of overrepresented sequences in 18 out of 24 samples. These sequences were not adapters and were the same in all samples as the repeat regions contained a few Cs and high number of Ts. Since the trimming would not take care of that and the aligning software would ignore them, the pipeline was continued without trimming the datasets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sample ID** | **GC (X)** | **TC (X)** | **Uniquely Aligned Reads (%)** | **Multi-mapped Reads (%)** | **Overall Alignment (%)** |
| Mouse 1 | 0.734 | 3262.4 | 81.76 | 9.01 | 95.37 |
| Mouse 2 | 0.799 | 3553.8 | 82.17 | 9.32 | 95.93 |
| Mouse 3 | 0.659 | 2934.2 | 81.56 | 9.88 | 96.01 |
| Mouse 4 | 0.819 | 3646.1 | 81.41 | 9.21 | 95.39 |
| Mouse 5 | 0.724 | 3219.5 | 81.49 | 8.98 | 95.58 |
| Mouse 6 | 0.971 | 4317.8 | 59.12 | 6.59 | 69.57 |
| Mouse 7 | 0.648 | 2883.7 | 80.5 | 10.28 | 95.47 |
| Mouse 8 | 0.625 | 2781.6 | 81.55 | 9.36 | 95.78 |
| Mouse 9 | 0.686 | 3051.6 | 80.29 | 9.07 | 95.04 |
| Mouse 10 | 0.654 | 2907.1 | 81.22 | 9.26 | 95.67 |
| Mouse 11 | 0.242 | 1078.3 | 79.06 | 10.53 | 94.37 |
| Mouse 12 | 0.66 | 2936.2 | 81.64 | 9.08 | 95.72 |

**Table 1. Summary of genomic coverage (GC) and transcriptome coverage (TC), reads (uniquely aligned and multimapped), and general alignment rates with mouse reference genome.**

*HISAT2 (v2.1.0), SAMtools(v0.1.19),* *Stringtie (v1.3.4), and gffcompare (v0.11.5)*

HISAT2 was utilized to align each sample individually to the mouse reference genome. At the conclusion of each sample run a printed message was sent to the “standard error” file handle summarizing results, this information was utilized to generate the remaining columns found in **Table 1**. The summary included number of reads, percentage of reads that paired, and concordant alignments. The general alignment rate, percentage of uniquely aligned reads, and percentage of multi-aligned reads for each sample is summarized in **Table 1**. Overall, each sample had an alignment rate of approxiamtely 95% excluding mouse 6, uniquely aligned reads for most samples were approximately 80%, except mouse number 6 (59.12), and multi-mapped reads for all samples were less less than 11% as shown in **Table 1**. *Samtools* were employed next to convert the .sam files to binary files (.bam) as well as to sort the found sequences by genomic position in reference genome. *Stringtie* (in three distinct steps) assembled the alignments into full and partial transcripts, creating multiple isoforms as necessary and estimated the expression levels of all genes and transcripts. Five files in .ctab format were generated for each sample, housing separately the information on exons, introns, and their respective transcriptomes. The merged gtf file obtained via the *stringtie* merge step of the pipeline was further analyzed via the *gffcompare* utility which enables the comparison of assembled transcripts with the reference annotation. The *gffcompare* command resulted in several files, one of which contained statistical information summarizing the transcript assembly. Overall, the merged dataset containing the transcripts from all 12 mice consisted of 142,847 transcripts within 46,160 loci. Of the loci, 5,353 were novel. Of 394,081 exons, 8,628 (2.2%) were classified as novel.

*Post Processing* (*Ballgown (v2.18.0))*

Output from the *Ballgown*, run in R environment (RStudio *v1.2.5019*), using published protocol (*9*) identified 1325 genes with p-value <0.05, and two genes with a q-value <0.05. HISAT2 was utilized to align each sample individually to the mouse reference genome. At the conclusion of each sample run a printed message was sent to the “standard error” file handle summarizing results, this information was utilized to generate the remaining columns found in **Table 1**. The summary included number of reads, percentage of reads that paired, and concordant alignments. The general alignment rate, percentage of uniquely aligned reads, and percentage of multi-aligned reads for each sample is summarized in **Table 1**. Overall, each sample had an alignment rate of approximately 95% excluding mouse 6, uniquely aligned reads for most samples were approximately 80%, except mouse number 6 (59.12), and multi-mapped reads for all samples were less than 11% as shown in **Table 1**. *Samtools* were employed next to convert the .sam files to binary files (.bam) as well as to sort the found sequences by genomic position in reference genome. *Stringtie* (in three distinct steps) assembled the alignments into full and partial transcripts, creating multiple isoforms as necessary and estimated the expression levels of all genes and transcripts. Five files in. ctab format were generated for each sample, housing separately the information on exons, introns, and their respective transcriptomes. The merged gtf file obtained via the stringtie merge step of the pipeline was further analyzed via the *gffcompare* utility which enables the comparison of assembled transcripts with the reference annotation. The *gffcompare* command resulted in several files, one of which contained statistical information summarizing the transcript assembly. Overall, the merged dataset containing the transcripts from all 12 mice consisted of 142,847 transcripts within 46,160 loci. Of the loci, 5,353 were novel. Of 394,081 exons, 8,628 (2.2%) were classified as novel.

A

B

C

D

E

F

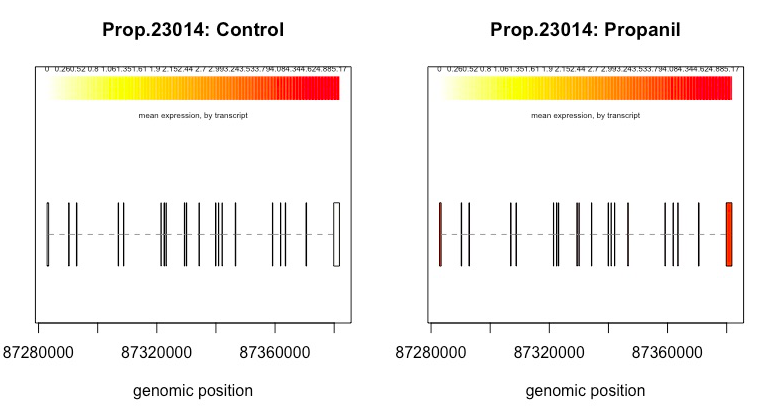
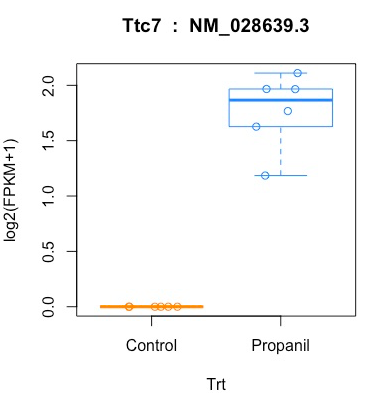
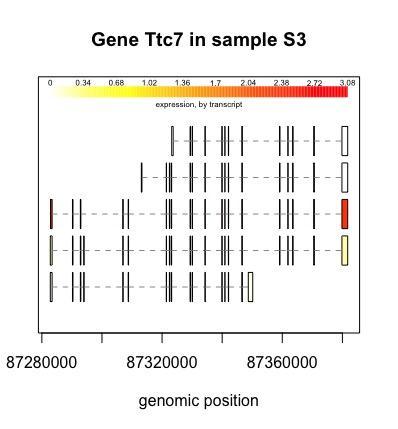
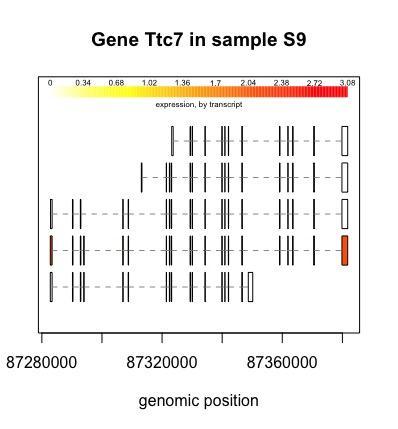
G

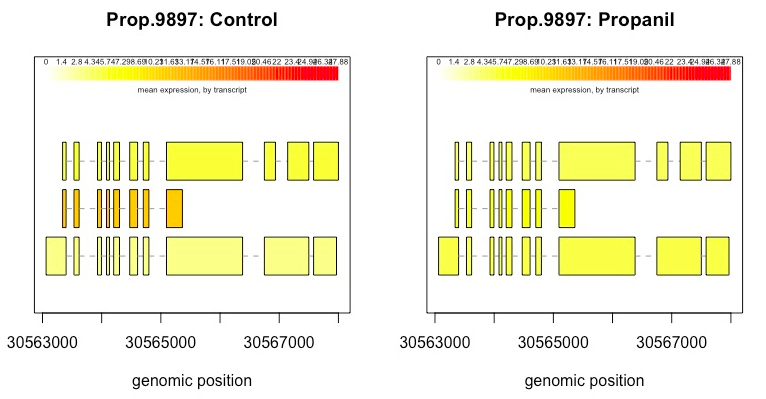
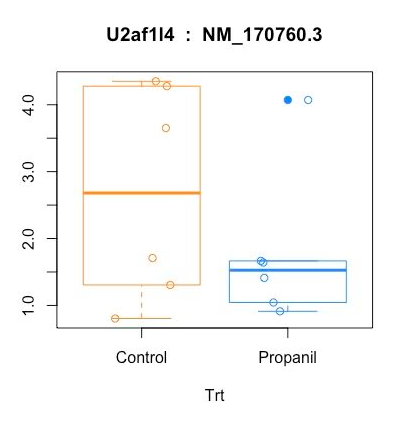
A

B

C

D



**Figure 1. Differential expression of genes and transcript in ovaries of propanil-treated mice.** Distributions of Log2 ofFPKM+1 in control (n=6) and propanil (n=6) mice for Ttc7 (**A**).Distributions of Log2 ofFPKM+1 in control (n=6) and propanil (n=6) mice for U2af1/4 (**B**). Representative images of the structure and expression levels of five distinct isoforms of the Ttc7 gene in a control (Sample 9; **C**) and propanil (Sample 3; **D**) – treated mice. Rows with the horizontal dash lines in the middle represent the transcripts found for the same gene. Short vertical lines and blocks represent the alignment of the transcripts with the reference genome, the width of the block indicates the length of the sequence and genomic position is indicated on the X axis. Expression levels are from the unfiltered output and are shown in varying shades from white to yellow to red, with red indicating a higher expression. After data filtering (omitting all transcripts with variance less than 1), only one transcript NM\_028639.3 of the five demonstrated significant difference between propanil and control mice (**E, F**). Mean expression of the Ttc7 gene in ovaries for all mice in control (**E**) and propanil (**F**) treated groups. Mean Ttc7 transcript expression was higher due to propanil treatment as indicated by darker shade of red. The gene, U2af1/4, was differentially expressed between propanil and control ovaries (**G, H**), however, three individual transcripts were not differentially expressed. Collectively, U2af1/4 was upregulated in propanil-treated mice (fold change 0.76, q=0.02).

D

F

G

H

E

C

B

A

H

E

F

G

H

**DISCUSSION**

The aim of this project was to utilize RNASeq data to determine differences in global gene expression of mice treated with propanil and HKSP as well as mice only challenged with HKSP (control). Based on the literature and preliminary data, the expectations were to find expression of genes that code for enzymes involved in the synthesis of reproductive hormones, pro- and anti-inflammatory cytokines, and tissue remodeling molecules to be modified by propanil treatment.

While assessing FastQC reports for all samples, it was evident that the same or similar repeated sequences consisting of a repeat of C’s followed by a long series of T’s were appearing for each sample. This was attributed to sequencing of mRNA poly A tails added through polyadenylation. The read alignments for each sample excluding sample 6 resulted in a high ( >95%) alignment rate, therefore, it was anticipated that the transcriptome assembler (*Stringtie*) would assemble the transcripts with high precision and sensitivity to aid in the detection of genes expressed at a low level. The percent of misaligned reads was minimal for all samples (<11%) which aided in the accuracy of assembly as misaligned reads can negatively influence the downstream assembly which, in return, could disrupt transcript and gene expression estimates. After inspection of the merged transcripts from all samples, *gffcompare* determined 11.6% of loci and 2.2% of exons were identified to be novel in relation to the reference genome. These values are relatively low as the mice utilized for the study were highly inbred and should contain little genetic variation. In the future, *gffcompare* could be used to look at transcript uniqueness between each sample and between treatment groups. Inspection of the FPKM distribution across all samples indicated the average Log2 FPKM was <1 which indicates a very small number of transcripts were expressed at high levels across all samples.

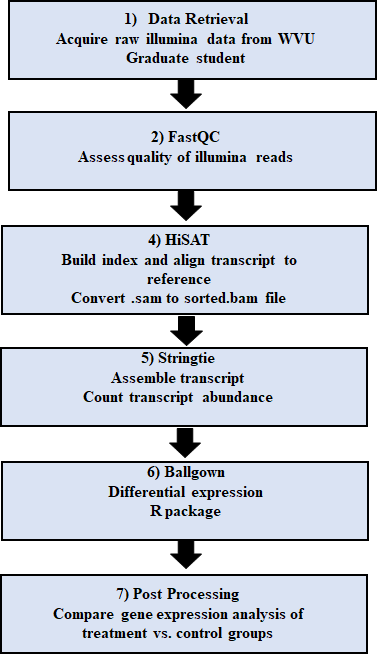
With the given settings in bioinformatics steps (*9*), (filtering away all samples with variance less than 1, q-value <0.05, etc.) we found two genes and one transcript being differentially expressed. While the number of detected genes and transcript were less than expected, the two genes found differentially expressed, are interesting. Both the transcript and its corresponding gene, tetratricopeptide Repeat Domain 7A (*Ttc7*) were overexpressed in ovaries of propanil-treated animals. There are 17 articles about gene function of *Ttc7* available on pubmed. Recent reports show that mutations of *Ttc7* gene lead to multisystemic problems, mostly affecting epithelial and immune tissues (*10*). Deficiency in *Ttc7* gene in both humans and mice, leads to early onset of inflammatory bowel disease, various degrees of immunodeficiency and very low level of lymphocytes in the blood (*11*). On the other hand, the over expression of *Ttc7* in the brain was shown to be linked to the plague formation in Alzheimer’s disease (*12*). The highest expression of *Ttc7* in was found in adult mouse duodenum (23 RPKM), in the spleen, small intestine and testes, with the mild expression found in ovaries (5 RPKM) (*13*). Functions of this gene include a role in iron transport and anemia. Deletion of the *Ttc7* was found in hereditary anemic mice (*14*). Knowing propanil leads to methemoglobinemia and anemia, we predict that ovarian tissue may react in defense by upregulation of *Ttc7* to restore oxygen homeostasis in the tissue (*3*, *15*).

Additional gene differentially expressed between propanil-treated and control ovaries was a gene encoding U2 small nuclear ribonucleoprotein auxiliary factor (*U2af1/4*). Lower expression was observed in propanil-treated mice than in controls (fold change 0.76). Pub med search lists 56 citations on *U2af* gene function. The highest expression of *U2af* in mice was found in embryonic (day 11.5) central nervous system (119 RPKM), then in the placenta (70 RPKM) and mild expression found in ovaries (~10 RPKM) and many other tissues (*13*). *U2af1* is a part of spliceosome machinery, responsible for splicing (and ligating) the coding regions (exons) from a pre-cursor RNA, while at the same time it cuts out the noncoding regions (introns) (*16*). A conditional knock-in allele of a mutant splicing factor gene *U2af1* led to impaired formation of blood cells and development of leukemia in mice (*17*). A whole transcriptome (RNA-seq) analysis of mice mutant in *U2AF1* gene demonstrated altered hematopoiesis and changes in pre-mRNA splicing in hematopoietic progenitor cells (*18*). In addition, leukopenia, a decrease in white blood cells was observed (*18*). The implications of this finding to our research scenario still needs to be determined. Since the estrous cycle in mice is on average four days, ovary undergoes fast tissue remodeling and shifts in hormonal production. It is known that ovulation involves the influx of inflammatory cytokines and chemokines, and immune cells. Additional results in our experiment detected a significant decrease in spleen size and number of white blood cells in the spleen in propanil-treated mice concomitantly with an increase in progesterone from the ovaries. Thus, decrease of the splicing factor *U2af1* may lead to altered cellular and functional make-up of the ovary. Further research will confirm hypotheses about the function of the differentially expressed genes and transcript.

Inspection of the FPKM distribution across all samples indicated the average Log2 FPKM was <1, which indicates a very small number of transcripts were expressed at high levels across all samples. Only two genes were found to be differentially expressed in ovaries of propanil-treated mice, which was below expectations. The low count could be due to various reasons—one of which could be the mere biology, meaning the propanil treatment does not lead to significant changes in ovaries in 24 hours. This may be a completely valid conclusion; however, two sets of preliminary data indicate otherwise. A microarray analysis of ovarian tissues, using the same experimental design and strain of mice, identified more than 100 differentially expressed RNAs, and in a real time RT-PCR were at least 3 differentially expressed genes, coding for steroidogenic enzymes, revealed. Although HISAT2 was the fastest splice-aware mapper out of 14 recently compared algorithms, HISAT2 also had a low recall percentage when mapping reads with high complexity, i.e. more polymorphic sites and higher error rates, on the default settings (*19*) and mapping accuracy was vastly improved after tuning the parameters. Therefore, tuning the parameters in the pipeline may allow for less robust estimations of expression. The parameters in post-processing may have also contributed to few differences in fold change as the variance cut-off value of 1 could be modified. If more stringent criteria are used, for example, using the variance of 1.5 or 2, larger number of genes/transcripts should be filtered out. The total number of genes/transcripts kept for statistical analysis would be smaller. Since the total number of genes used for statistical tests is factored into the q-values calculation and the number is significant genes based on the p-value should stay the same, the final results, as for number of genes/transcripts meeting the q-values cut-off may change.

In summary, two genes were found to be differentially expressed in the ovaries of propanil-treated mice. One challenge was lack of knowledge on R software. We remained stuck for several hours on step when Ballgown in R would not read the data correctly and we kept getting error message about the missing lines. The culprit was the last sample of 12, from which the files from scratch folder of Spruce were copied wrong way as empty files. Order of samples metadata (original order: mouse 1, mouse 2, …, mouse 12, with corresponding treatment assignment) provided to R is also very important, as the *ballgown* (R) will be taking the files in the computer-preferred-kind of order, such as mouse 1, mouse 11, mouse 12, 3, 4, 5, …9). Thus, the order statement was utilized [not in the Nature paper (*9*)] so the order of phenotypic information matched the order of aligned files utilized by R. Staying organized with folders and subfolders, both inside the Spruce and in working environment for R proved to be crucial in preventing over-writing of results and loss of earlier outputs, especially in *stringtie count* step.

**PIPELINE CODE**

**

**Figure 2: An overview of the pipeline for RNA sequencing analysis**

***Fastqc (v0.11.7)***

fastqc S5\_R5.fastq.gz

#fastqc was utilized to address the quality of the raw paired end dataset selected

***HISAT2 (v2.1.0), SAMtools (v0.1.19)* and *Stringtie (v1.3.4)***

hisat2-build mouseref.fa Mouse

#reference genome was indexed for alignment

hisat2 -p 8 -x Mouse -1 S5\_R1.fastq.gz -2 S5\_R2.fastq.gz -S /scratch/demiddleton/Propanil/S5-aligned.sam

#each sample was read individually for alignment

samtools view -Su S5-aligned.sam | samtools sort - S5.sorted

#conversion of files, compact files to save time and space and sorting in order of reference genome

stringtie -p 5 -G mouseref.gtf -o S5-transcripts.gtf -l Prop-5 S5.sorted.bam

#assembly of RNA-sequences alignments

stringtie --merge -p 8 -G mouseref.gtf -o stringtie\_merged.gtf Prop\_mergelist.txt

#counting number of transcripts per sample

stringtie -e -B -p 1 -G stringtie\_merged.gtf -o /scratch/demiddleton/Propanil/MCounts/S5/S5-count.gtf S5.sorted.bam

#Count transcript abundances per sample

***gffcompare***

gffcompare -G -r mouseref.gtf stringtie\_merged.gtf

#compare gtf file produced by stringtie with the mouse reference annotation file. The -r indicates the reference genome and -G tells gffcompare to compare all transcripts in the input file.

***Ballgown (v2.18.0) in R*** *(RStudio v1.2.5019)*

>pheno\_data = read.csv("propanil.csv")

#Load phenotype data for the samples

pheno\_data = pheno\_data[order(pheno\_data$Ids),]

#ballgown (or R) will order of the pheno\_data in a computer format 1, 10, 11, 12, 2, 3, 4, .....

>PropanilResults = ballgown(dataDir = "MCounts", samplePattern = "M", pData=pheno\_data) #Read expression data that were calculated by Stringtie. Here, the data directory dataDir was MCounts and the pattern that appears in the sample names (samplePattern) was M. The phenotypic information that we loaded in the previous step (pData) was called pheno\_data.

>PropanilResults\_filt = subset(PropanilResults,"rowVars(texpr(PropanilResults)) >1",genomesubset=TRUE)

#Filter to remove low-abundance genes

>results\_transcripts = stattest(PropanilResults\_filt, feature="transcript",covariate="Treatment", getFC=TRUE, meas="FPKM")

#Identify *transcripts* that show statistically significant differences between groups.

>results\_genes = stattest(PropanilResults\_filt, feature="gene", covariate="Treatment", getFC=TRUE, meas="FPKM")

#Identify genes that show statistically significant differences between groups.

>results\_transcripts = data.frame(geneNames=ballgown::geneNames(PropanilResults\_filt), geneIDs=ballgown::geneIDs(PropanilResults\_filt), results\_transcripts)

#Add gene names and gene IDs to the results\_transcripts data frame:

>results\_transcripts = arrange(results\_transcripts,pval)

>results\_genes = arrange(results\_genes,pval)

#Sort the results from the smallest P value to the largest:

>write.csv(results\_transcripts, "Propanil\_transcript\_results.csv", row.names=FALSE)

>write.csv(results\_genes, "Propanil\_gene\_results.csv", row.names=FALSE)

#Write the results to a csv file that can be shared and distributed

>subset(results\_transcripts,results\_transcripts$qval<0.05)

geneNames geneIDs feature id fc pval qval

1 Ttc7 Prop.23014 transcript 124813 3.455129 1.076726e-06 0.03218441

>subset(results\_genes,results\_genes$qval<0.05)

#Identify transcripts and genes with a q value <0.05

> subset(results\_genes,results\_genes$qval<0.05)

feature id fc pval qval

1 gene Prop.23014 3.4591467 7.087585e-07 0.009283319

2 gene Prop.9897 0.7615516 2.475814e-06 0.016214109

#graphs and plots (steps 17-21 from reference 9):

> library(RSkittleBrewer)

> tropical= c('darkorange', 'dodgerblue','hotpink', 'limegreen', 'yellow')

> palette(tropical)

> fpkm = texpr(PropanilResults,meas="FPKM")

> fpkm = log2(fpkm+1)

> boxplot(fpkm,col=as.numeric(pheno\_data$Trt),las=2,ylab='log2(FPKM+1)')

> ballgown::transcriptNames(PropanilResults)[1] ##1 ## "Prop.23014"

> ballgown::transcriptNames(PropanilResults)[1] ##1 ## "Prop.23014"

> ballgown::geneNames(PropanilResults) [1] ## "Ttc7

> ballgown::transcriptNames(PropanilResults)[124813]

> ballgown::geneNames(PropanilResults) [124813]

> plot(fpkm[124813,] ~ pheno\_data$Trt, border=c(1,2), main=paste(ballgown::geneNames(PropanilResults)[124813],' : ',ballgown::transcriptNames(PropanilResults)[124813]),pch=19, xlab="Trt",ylab='log2(FPKM+1)')

> points(fpkm[124813, ] ~ jitter(as.numeric(pheno\_data$Trt)), col=as.numeric(pheno\_data$Trt))

> boxplot(fpkm,col=as.numeric(pheno\_data$Trt),las=2,ylab='log2(FPKM+1)')

> plotTranscripts(ballgown::geneIDs(PropanilResults)[124813], PropanilResults, main=c('Gene Ttc7 in

+ sample S1'), sample=c('S1'))

> plotTranscripts(ballgown::geneIDs(PropanilResults)[124813], PropanilResults, main=c('Gene Ttc7 in sample S1'), sample=c('S1'))

> plotTranscripts(ballgown::geneIDs(PropanilResults)[124813], PropanilResults, main=c('Gene Ttc7 in sample S2'), sample=c('S2'))

> plotTranscripts(ballgown::geneIDs(PropanilResults)[124813], PropanilResults, main=c('Gene Ttc7 in sample S3'), sample=c('S3'))

> plotTranscripts(ballgown::geneIDs(PropanilResults)[124813], PropanilResults, main=c('Gene Ttc7 in sample S4'), sample=c('S4'))

> plotTranscripts(ballgown::geneIDs(PropanilResults)[124813], PropanilResults, main=c('Gene Ttc7 in sample S5'), sample=c('S5'))

> plotTranscripts(ballgown::geneIDs(PropanilResults)[124813], PropanilResults, main=c('Gene Ttc7 in sample S6'), sample=c('S6'))

> plotTranscripts(ballgown::geneIDs(PropanilResults)[124813], PropanilResults, main=c('Gene Ttc7 in sample S7'), sample=c('S7'))

> plotTranscripts(ballgown::geneIDs(PropanilResults)[124813], PropanilResults, main=c('Gene Ttc7 in sample S8'), sample=c('S8'))

> plotTranscripts(ballgown::geneIDs(PropanilResults)[124813], PropanilResults, main=c('Gene Ttc7 in sample S9'), sample=c('S9'))

> plotTranscripts(ballgown::geneIDs(PropanilResults)[124813], PropanilResults, main=c('Gene Ttc7 in sample S10'), sample=c('S10'))

> plotTranscripts(ballgown::geneIDs(PropanilResults)[124813], PropanilResults, main=c('Gene Ttc7 in sample S11'), sample=c('S11'))

> plotTranscripts(ballgown::geneIDs(PropanilResults)[124813], PropanilResults, main=c('Gene Ttc7 in sample S12'), sample=c('S12'))

> plotMeans(ballgown::geneIDs(PropanilResults)[9897], PropanilResults\_filt, groupvar="Trt",legend=TRUE)

> plotMeans(ballgown::geneIDs(PropanilResults)[54030], PropanilResults\_filt, groupvar="Trt",legend=TRUE)

> plotMeans(ballgown::geneIDs(PropanilResults)[54031], PropanilResults\_filt, groupvar="Trt",legend=TRUE)

> plotMeans(ballgown::geneIDs(PropanilResults)[54032], PropanilResults\_filt, groupvar="Trt",legend=TRUE)

> ballgown::transcriptNames(PropanilResults)[54030]

> ballgown::transcriptNames(PropanilResults)[54031]

> ballgown::transcriptNames(PropanilResults)[54032]

> plot(fpkm[54032,] ~ pheno\_data$Trt, border=c(1,2), main=paste(ballgown::geneNames(PropanilResults)[54032],' : ',ballgown::transcriptNames(PropanilResults)[54032]),pch=19, xlab="Trt",ylab='log2(FPKM+1)')

> points(fpkm[54032, ] ~ jitter(as.numeric(pheno\_data$Trt)), col=as.numeric(pheno\_data$Trt))

> plot(fpkm[54031,] ~ pheno\_data$Trt, border=c(1,2), main=paste(ballgown::geneNames(PropanilResults)[54031],' : ',ballgown::transcriptNames(PropanilResults)[54031]),pch=19, xlab="Trt",ylab='log2(FPKM+1)')

> points(fpkm[54031, ] ~ jitter(as.numeric(pheno\_data$Trt)), col=as.numeric(pheno\_data$Trt))

> plot(fpkm[54030,] ~ pheno\_data$Trt, border=c(1,2), main=paste(ballgown::geneNames(PropanilResults)[54030],' : ',ballgown::transcriptNames(PropanilResults)[54030]),pch=19, xlab="Trt",ylab='log2(FPKM+1)')

> points(fpkm[54030, ] ~ jitter(as.numeric(pheno\_data$Trt)), col=as.numeric(pheno\_data$Trt))

**REFERENCES**

1. R. Bartha, D. Pramer, in *Advances in Applied Microbiology*, D. Perlman, Ed. (Academic Press, 1970; http://www.sciencedirect.com/science/article/pii/S0065216408704088), vol. 13, pp. 317–341.

2. R. E. Hoagland, J. K. Norsworthy, F. Carey, R. E. Talbert, Metabolically based resistance to the herbicide propanil in Echinochloa species. *Weed Science*. **52**, 475–486 (2004).

3. D. C. McMillan, T. A. McRae, J. A. Hinson, Propanil-induced methemoglobinemia and hemoglobin binding in the rat. *Toxicol. Appl. Pharmacol.* **105**, 503–507 (1990).

4. J. B. Barnett, J. Gandy, Effect of acute propanil exposure on the immune response of C57BI6 mice. *Fundamental and Applied Toxicology*. **12**, 757–764 (1989).

5. M. A. Valentovic, T. Yahia, J. G. Ball, S. K. Hong, P. I. Brown, G. O. Rankin, 3,4-Dicholoroaniline acute toxicity in male Fischer 344 rats. *Toxicology*. **124**, 125–134 (1997).

6. M. Eddleston, M. Rajapakshe, D. Roberts, K. Reginald, M. H. R. Sheriff, W. Dissanayake, N. Buckley, Severe Propanil [N-(3,4-dichlorophenyl) propanamide] Pesticide Self-Poisoning. *J Toxicol Clin Toxicol*. **40**, 847–854 (2002).

7. K. D. Salazar, M. R. Miller, J. B. Barnett, R. Schafer, Evidence for a Novel Endocrine Disruptor: The Pesticide Propanil Requires the Ovaries and Steroid Synthesis to Enhance Humoral Immunity. *Toxicol Sci*. **93**, 62–74 (2006).

8. Z.-Q. Wu, Q. Vos, Y. Shen, A. Lees, S. R. Wilson, D. E. Briles, W. C. Gause, J. J. Mond, C. M. Snapper, In Vivo Polysaccharide-Specific IgG Isotype Responses to Intact Streptococcus pneumoniae Are T Cell Dependent and Require CD40- and B7-Ligand Interactions. *The Journal of Immunology*. **163**, 659–667 (1999).

9. M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, S. L. Salzberg, Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*. **11**, 1650–1667 (2016).

10. U. Nüesch, A. A. Mauracher, L. Opitz, B. Volkmer, K. Michalak-Mićka, J. Kamarashev, T. Hartwig, E. Reichmann, B. Becher, S. Vavassori, J. Pachlopnik Schmid, Epithelial proliferation in inflammatory skin disease is regulated by tetratricopeptide repeat domain 7 (Ttc7) in fibroblasts and lymphocytes. *J. Allergy Clin. Immunol.* **143**, 292-304.e8 (2019).

11. M.-T. El-Daher, J. Lemale, J. Bruneau, C. Leveau, F. Guerin, N. Lambert, J.-S. Diana, B. Neven, F. E. Sepulveda, A. Coulomb-L’Hermine, T. Molina, C. Picard, A. Fischer, G. de Saint Basile, Chronic Intestinal Pseudo-Obstruction and Lymphoproliferative Syndrome as a Novel Phenotype Associated With Tetratricopeptide Repeat Domain 7A Deficiency. *Front Immunol*. **10**, 2592 (2019).

12. M. Sun, Y. Zhao, M. Han, B. Zhang, X. Zhang, Q. Zhang, N. K.-H. Lim, W.-A. Wang, F.-D. Huang, TTC7 and Hyccin Regulate Neuronal Aβ42 Accumulation and its Associated Neural Deficits in Aβ42-Expressing Drosophila. *J. Alzheimers Dis.* **65**, 1001–1010 (2018).

13. F. Yue, Y. Cheng, A. Breschi, J. Vierstra, W. Wu, T. Ryba, R. Sandstrom, Z. Ma, C. Davis, B. D. Pope, Y. Shen, D. D. Pervouchine, S. Djebali, R. E. Thurman, R. Kaul, E. Rynes, A. Kirilusha, G. K. Marinov, B. A. Williams, D. Trout, H. Amrhein, K. Fisher-Aylor, I. Antoshechkin, G. DeSalvo, L.-H. See, M. Fastuca, J. Drenkow, C. Zaleski, A. Dobin, P. Prieto, J. Lagarde, G. Bussotti, A. Tanzer, O. Denas, K. Li, M. A. Bender, M. Zhang, R. Byron, M. T. Groudine, D. McCleary, L. Pham, Z. Ye, S. Kuan, L. Edsall, Y.-C. Wu, M. D. Rasmussen, M. S. Bansal, M. Kellis, C. A. Keller, C. S. Morrissey, T. Mishra, D. Jain, N. Dogan, R. S. Harris, P. Cayting, T. Kawli, A. P. Boyle, G. Euskirchen, A. Kundaje, S. Lin, Y. Lin, C. Jansen, V. S. Malladi, M. S. Cline, D. T. Erickson, V. M. Kirkup, K. Learned, C. A. Sloan, K. R. Rosenbloom, B. Lacerda de Sousa, K. Beal, M. Pignatelli, P. Flicek, J. Lian, T. Kahveci, D. Lee, W. J. Kent, M. Ramalho Santos, J. Herrero, C. Notredame, A. Johnson, S. Vong, K. Lee, D. Bates, F. Neri, M. Diegel, T. Canfield, P. J. Sabo, M. S. Wilken, T. A. Reh, E. Giste, A. Shafer, T. Kutyavin, E. Haugen, D. Dunn, A. P. Reynolds, S. Neph, R. Humbert, R. S. Hansen, M. De Bruijn, L. Selleri, A. Rudensky, S. Josefowicz, R. Samstein, E. E. Eichler, S. H. Orkin, D. Levasseur, T. Papayannopoulou, K.-H. Chang, A. Skoultchi, S. Gosh, C. Disteche, P. Treuting, Y. Wang, M. J. Weiss, G. A. Blobel, X. Cao, S. Zhong, T. Wang, P. J. Good, R. F. Lowdon, L. B. Adams, X.-Q. Zhou, M. J. Pazin, E. A. Feingold, B. Wold, J. Taylor, A. Mortazavi, S. M. Weissman, J. A. Stamatoyannopoulos, M. P. Snyder, R. Guigo, T. R. Gingeras, D. M. Gilbert, R. C. Hardison, M. A. Beer, B. Ren, Mouse ENCODE Consortium, A comparative encyclopedia of DNA elements in the mouse genome. *Nature*. **515**, 355–364 (2014).

14. R. A. White, S. G. McNulty, N. N. Nsumu, L. A. Boydston, B. P. Brewer, K. Shimizu, Positional cloning of the Ttc7 gene required for normal iron homeostasis and mutated in hea and fsn anemia mice. *Genomics*. **85**, 330–337 (2005).

15. D. C. McMillan, T. P. Bradshaw, J. A. Hinson, D. J. Jollow, Role of metabolites in propanil-induced hemolytic anemia. *Toxicol. Appl. Pharmacol.* **110**, 70–78 (1991).

16. Spliceosome - an overview | ScienceDirect Topics, (available at https://www.sciencedirect.com/topics/medicine-and-dentistry/spliceosome).

17. D. L. Fei, T. Zhen, B. Durham, J. Ferrarone, T. Zhang, L. Garrett, A. Yoshimi, O. Abdel-Wahab, R. K. Bradley, P. Liu, H. Varmus, Impaired hematopoiesis and leukemia development in mice with a conditional knock-in allele of a mutant splicing factor gene U2af1. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E10437–E10446 (2018).

18. C. L. Shirai, J. N. Ley, B. S. White, S. Kim, J. Tibbitts, J. Shao, M. Ndonwi, B. Wadugu, E. J. Duncavage, T. Okeyo-Owuor, T. Liu, M. Griffith, S. McGrath, V. Magrini, R. S. Fulton, C. Fronick, M. O’Laughlin, T. A. Graubert, M. J. Walter, Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo. *Cancer Cell*. **27**, 631–643 (2015).

19. G. Baruzzo, K. E. Hayer, E. J. Kim, B. Di Camillo, G. A. FitzGerald, G. R. Grant, Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Methods*. **14**, 135–139 (2017).