**PROJECT 3 Proposal**

**Group: nkob**

**Members: Heather Baldwin, Ida Holásková, Denzel Middleton**

**Date November 19, 2019**

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

**BACKGROUND**

     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 (*4*–*7*). 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 (*8*). 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 (*8*, *9*). 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 was designed, in which female mice 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 is 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). The innovative components of this experiment are the use of cutting-edge technology (Illumina sequencing and data processing with the newest available software) to analyze the ovarian transcriptome of propanil-exposed mice as well as the timing of the tissue collection (24 hours). Previous research focused on effects due to propanil seven days after exposure. To date, ovarian RNA Seq data post propanil challenge are not publicly available.

**METHODS**

**Data Availability and Retrieval**

To test the experimental hypothesis that ovarian gene expression is modified after propanil exposure, paired end Illumina sequencing data will be 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),six mice were treated with propanil and six were treated with peanut oil (control). Raw data in the .fastq file format will be downloaded from the sequencing core web page. Raw reads will be mapped to the reference genome (Accession: GCA\_000001635.8), which is available from NCBI’s assembly platform in the .fna, .gff file formats. This reference genome assembly has 2,730,855,475 bp with 605 contigs and 162 scaffolds.

**Experimental Software Pipeline**

The workflow of raw data processing will be carried out via the High-Performance Computing center (Spruce Knob) and the use of R *(v3.6.1)*, available at West Virginia University. The experimental software pipeline is demonstrated in Figure 1 (*10*).

*Fastqc (v0.11.7)* and *Trimmomatic (v0.38)*

 To evaluate the quality of raw datasets, FastQC will be used to assess the quality of the raw data as FastQC is both available on Spruce and is a widely accepted application for quality control (*11*). The resulting FastQC file will report the per base sequence quality, sequence content, N content, sequence duplication levels, overrepresented sequences, and adapter content of the dataset. Coverage of each of the twelve samples will be calculated to check if the data contains reads with an abnormally low or high coverage. Transcriptome size of the mouse genome will give us a useful guideline on the coverage. Trimmomatic will be used to trim the dataset as it is specifically designed for paired-end data and available on Spruce (*12*). If preliminary Fastqc report exhibit that reads decline in quality (Q < 30), Trimmomatic will be utilized to trim the low quality reads.

*HISAT2 (v2.1.0)*

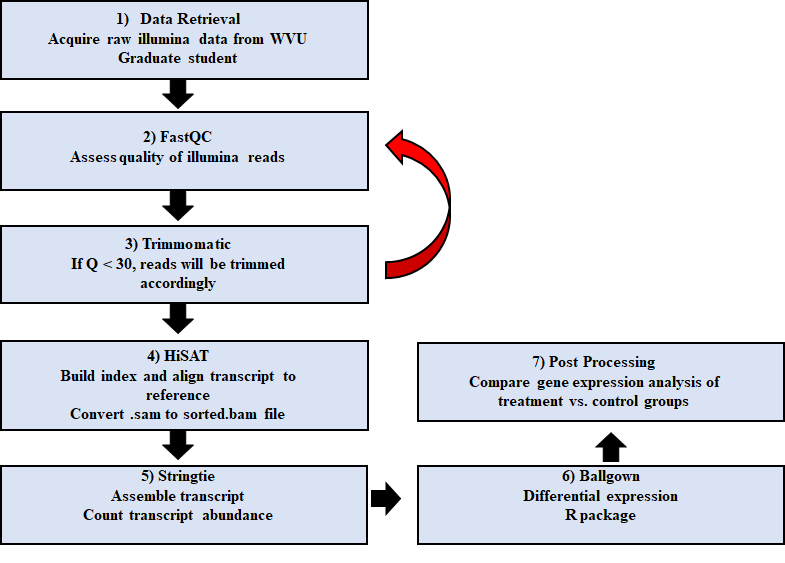
Following quality control and trimming, HISAT2 alignment program for mapping next-generation Illumina sequence reads to the C57Bl/6J reference genome. HISAT2 is a fast and sensitive spliced alignment program for mapping next generation sequencing reads, the program uses a large set of small graph FM index (GFM) indexes that collectively cover the whole genome. Spliced junction mapping is critical for mapping reads across splice junctions when using eukaryotic data. The use of Ferragina-Manzini (FM) index allows the rapid search of genomes, yielding speeds of millions of reads per hour. The reference genome file format will be converted from .fna to .fa in scratch before beginning the alignment. HISATRef is the first script in which indexes are built for the alignment. The second script is HISATAlign, in which the alignment of the illumina reads to the reference genome will occur, from this .sam files will be created. Using the SamSort command the .sam files will be converted to a sorted.bam files.

*Stringtie (v1.3.4)*

The HISAT2 sorted.bam file output will be used for the next step, Stringtie. Stringtie is a software program that is fast and highly efficient at assembling RNA sequencing alignments into transcripts. Gene features without explicit exons are no longer taken as single-exon transcripts with this version. In order to begin assembly, stringtie will be run on every sample with the .gff annotated reference genome to generate a .gtf file. Stringtie merge will be the next command in which a list of transcripts are generated. The final step is stringtie count in which the transcript abundance is given. It will be important to make sure samples are run individually and in separate directories, in order to receive the adequate transcript quantification for each sample. Stringtie culminates the final step of the pipeline available on Spruce.

*Ballgown (v2.18.0)*

The stringtie output is a linked set of tables that can be directly read in R *(v3.6.1)* through functions in the Ballgown package. The tabled information includes phenotype data, expression data, and genomic information. Ballgown provides functions such as data visualization and inspection, statistical tests for differential expression, multiple test correction, and downstream inspection and summarization of results. Using the ballgown stattest command, a data table containing the transcripts ID and corresponding p- and q-values will be generated. Additionally, the ≥tFC option in stattest will be used to get the estimated fold change.

**

**Figure 1.** An overview of RNA sequencing protocol.

**Division of Labor**

Data retrieval from WVU Animal & Nutritional Sciences graduate student and reference genome collection from NCBI was completed by both IH and HB. Quality control and Trimmomatic (if necessary) on raw reads will be conducted by IH. Our 12 samples will be evenly divided amongst nkob members (4 per member) starting at HISAT2 and ending with post processing. We will work as a group to work through the pipeline with our first samples in order to assure that our commands are correct and running.The writing of this proposal was a combined group effort as the manuscript was broken into sections for each group member to complete. Issues throughout the process of implementation will be resolved by the use of software manuals or codes found online to assist in the completion of experimentation, if the issues are still difficult and unresolved we will then address them with Dr. Driscoll. Generation of the final report will be conducted by all nkob members with equal contribution to the writing of the manuscript.

**EXPECTED OUTCOME**

Through the utilization of RNA Seq data, ovarian transcriptome differential expression of C57Bl/6J mice will be determined in response to propanil/HKSP challenge in comparison to mice challenged solely with HKSP (control). After assessing the quality and trimming if necessary, the RNA Seq data will be indexed and aligned to the mouse reference genome. The transcripts will then be assembled and abundance will be noted. Further, we expect to obtain a list of all transcripts that were differentially expressed in relation to the control animals with associated log fold change (LogFC) values. A table with the top 10 up- and down-regulated genes will be generated and, if time permits, a pathway analysis will be conducted on genes up- and down-regulated. Specifically, a Gene Ontology (GO) analysis of statistically significant transcripts generated by the RNA-Seq data could be performed utilizing DAVID Bioinformatics Resource 6.8 (*13*). This may provide insight into the physiological effects of propanil on the mouse ovarian transcriptome. To ensure the correct labeling of files and to visualize the effect of propanil treatment, a principal component analysis (PCA) plot will be constructed using statistical software JMP (v. Pro 14). Throughout our proposed pipeline, several errors have the possibility of arising. After ensuring the reads are of good quality, reads will be aligned to the reference genome and then mapping statistics will be generated using SAMtools to address mapping quality. When looking at the quality of the mapping, the following statistics will be addressed: percentage of reads aligned, percentage of reads that have multiple alignments, percentage of reads that have a unique alignment, and the percentage of reads that are properly paired. If any of the previously listed statistics are unsatisfactory, troubleshooting will be conducted and may include ensuring the quality of the reference genome and/or altering alignment parameters. To prevent issues related to keeping files separate and unwanted merging of data, files will be carefully named and separately ran for specific steps of the pipeline. In addition, nkob group members will begin implementing the pipeline immediately to provide time for Spruce to process the data, possible troubleshooting, and repetition.

**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. 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).

7. 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).

8. 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).

9. 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).

10. 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).

11. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data, (available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

12. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 2114–2120 (2014).

13. D. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. **4**, 44–57 (2009).