Kelli McKeegan

Independent Project – Final Report

11 May 2021

**INTRODUCTION TO RESEARCH**

The Junonia coenia densovirus (JcDV) infects multiple species of Lepidoptera, being contracted orally and hindering larval respiration and molting (***1***). For my research, I study variation in the effects of JcDV across populations of a western North American Lycaenid that has only recently been discovered as a permissible host for JcDV: the Melissa blue butterfly (*Lycaeides melissa*). *L. melissa* undergoes 3-4 generations per season (May-September) and is non-migratory, overwintering in their patchy populations as eggs. Not much is known about transmission of JcDV, or even where it is in the wild and in what amount. Although it is nonenveloped, JcDV is a member of the Parvoviridae family, members of which known to be environmentally robust and able to survive in an active state for a period of time outside of the host (***2,3***). However, all host-pathogen relationships are affected by various abiotic (e.g., UV, temperature, wind) and biotic (e.g., tri-trophic interactions, carriers of infection) factors (***4,5***).

In order to begin understanding causes for variation in disease effects as well as routes of JcDV dispersal, one of my projects involves sampling from the surrounding environment in *L. melissa* communities to screen for JcDV with qPCR. I intend to identify the patterns of viral prevalence and load in the context of environmental exposure and ultimately paint the picture of where JcDV is across the landscape of these localities. I predict that *L. melissa* populations with high disease effects will reflect heavy viral presence on the surface of surrounding plants, in the soil, and other arthropods in the community – while populations with low effects might be more naïve to the virus (i.e., have little environmental JcDV exposure). Furthermore, in order to uncover potential temporal cycling, I will compare viral abundance across generations of this multivoltine host. This means that 4 of the 26 sites (ones close to my home base for convenience) was repeatedly sampled throughout the flying season for time series analyses. My prediction for this part is that JcDV abundance in *L. melissa* populations will be low in in the early-season generation (hypothesizing that individuals who “made it” through the winter were resistant to or negative for the virus), and increase with subsequent generations as host populations become more dense and the virus is transmitted.

**DATA & PROBLEM**

Currently this study is between its two stages: Field samples from each site (some being repeated) are collected, but awaiting laboratory qPCR screening. As could likely be assumed, the dataset formed is rather expansive and while I was the only one contributing to it, somewhat messy. There are over 20 columns, inconsistent site labels, a subset of sites that were sampled multiple times, and the data itself has a lot of fluctuation. Due to various climate variables (e.g., summer heat, wind, cloud cover), combined with the multivoltine life history of *L. melissa* (with varying density across generations unique to their habitat), there are some sites and days without data for a certain variable, and some I had to return to because my first attempt was a “bust” or too early. Moreover, a few more columns of data will be added once I run the samples through qPCR: Ct values for two replicates each, along with a column for mean and calculated load from a standard curve regression equation. The problem to be solved lies within the messy, data and numerous columns that need to be cleaned and easily accessible for appending and analyzing.

**GOALS**

The goals of my independent project were as follows:

* Read in and clean Excel data sheet
* Easily call numbers or determine how many samples I have of each variable collected
* Add columns once qPCR screening is complete, with math (mean and standard curve)
* Ultimately: Use the tools I have learned in Data Science for Biology to reorganize and manipulate data for convenience of future analyses.

**PROPOSED METHODS**

In order to achieve the aforementioned goals, I first proposed to clean the data with OpenRefine so that it is consistent, properly formatted, and unnecessary columns or NAs are dealt with appropriately. I also proposed using a few Python commands to summarize database features or to calculate data in columns such as mean and minimum/maximum viral frequency or load. I then wanted to build a relational database from the spreadsheet so that I could easily access specific data for certain variables or sites through queries. I wanted to find a way to keep my data organized for upcoming collections, and have it set up for appending with qPCR data once I complete that step.

**RESULTS**

Before anything, I made new tabs in my spreadsheet for version control and metadata, to keep things clear and organized as I made changes. I then loaded my very “wide” dataset into OpenRefine and started a project with it to check for and edit any inconsistencies. This program also helped me to count how many unique populations were sampled, how many samples were collected at each site, and other valuable count data (instead of Python commands as I had proposed). As a final step in OpenRefine, I made a new column and reformatted the roman numeral dates manually into ISO format. Next, I thought about my data in a relational context and reorganized the spreadsheet into tabs that made sense to the study: One for data related to each collection event, one for site/population data, a community pool table with the numbers of each type of sample collected at each site, and tabs for each group of qPCR data once it is collected (*Lycaeides melissa* adults and larvae, plant rinses and extractions and soil, and other arthropods). See the attached chart on the next page for the basic structure of the relational database. Each of these tabs were exported as a .csv file and imported into SQLite as a table for subsequent analyses and queries.

**DISCUSSION**

Overall, I learned a lot about version control and data management, as well as how to use OpenRefine and SQLite. I previously did not know about the existence of relational databases which will be such valuable skills for future collections. Understanding the relationships between certain aspects of complex data will contribute to my ability to record data in a much more organized, cleaner way. Once I obtain qPCR data, I will be able to input it into the tables in SQLite to calculate mean and viral load and conduct statistical analyses. Finally, learning how to upload our assignments to GitHub repositories is a unique skill that I will continue to use throughout my research career because of its awesome capabilities. Overall, this course gave me the experience and computational tools to help build and share good data over long-term community projects.

**Diagram

Description automatically generated with low confidence**

REFERENCES

1. Mutuel D, Ravallec M, Chabi B, Multeau C, Salmon JM, Fournier P, Ogliastro M. 2010. Pathogenesis of *Junonia coenia* densovirus in *Spodoptera frugiperda*: A route of infection that leads to hypoxia. *Virology*. 403(2):137-144.
2. Bergoin M and Tijssen P. 2010. Densoviruses: a highly diverse group of arthropod parvoviruses. *Insect Virology.* p. 57-90.
3. Williams, T. 2018. Viruses. In: Hajek N, Shapiro-Ilan D, editors. *Ecology of Invertebrate Diseases*. Hoboken, NJ: John Wiley & Sons. p. 213-285.
4. Ment D, Shikano I, Glazer I. 2018. Abiotic Factors. In: Hajek N, Shapiro-Ilan D, editors. *Ecology of Invertebrate Diseases*. Hoboken, NJ: John Wiley & Sons. p.143-169.
5. Cory JS, Deschodt PS. 2018. The Biotic Environment. In: Hajek N, Shapiro-Ilan D, editors. *Ecology of Invertebrate Diseases*. Hoboken, NJ: John Wiley & Sons. p. 187-212.