Parameter optimization of EthoVision XT® for automated quantification of spontaneous behaviors in *Drosophila melanogaster*

Presented to the faculty of Lycoming College in partial fulfillment of the requirements for Departmental Honors in Biology

By

Sohini Mukherjee

Lycoming College

May 2022

Approved by:

Dr. Mary Morrison

Dr. Holly Bendorf

Dr. Jeremy Ramsey

Dr. David Andrew

This work was supported by a Joanne and Arthur Haberberger Fellowship, awarded to Sohini Mukherjee at Lycoming College, Williamsport, PA.

Table of Contents

Abstract

Grooming is a quantitative behavior that is controlled by multiple genes and contributes to fitness in many animals including insects. Excessive grooming is demonstrated by animals that model human genetic intellectual disabilities, such as a loss of function mutation of the *Drosophila* ortholog of the human Fragile X Syndrome gene. Usually, manual scoring techniques are used for quantifying grooming and they rely on direct human observation of prerecorded videos of fly behavior. It is a time-consuming process and potentially subject to scorer bias. Many groups have developed automated analysis techniques to increase scoring throughput and consistency, but most of these approaches rely on very specific hardware and in-house computational expertise that render them difficult to replicate in other laboratories. We have been developing a semi-automated method that relies on commercially available software EthoVisionXT® software (Noldus Information Technology) and standard laboratory hardware. In this study, our goal was to identify the optimal parameter settings of EthoVisionXT® to generate grooming scores most congruent with a manually-scored set of benchmark videos. Our results show that an initial activity setting of 16, a movement threshold of 0.05 cm/s and activity thresholds of 10% (0.10), 14% (0.14), and 15% (0.15) yield the best set of automated scores that are comparable to manual scores. We also identified an optimization conflict in the comparison of automated and manual scores on a fly-by-fly basis versus at the population level, with different settings providing conflicting best results. Future studies should focus on working within these recommended software parameters, first with the same benchmark population and then with different populations of flies to finalize a software setting for automatic quantification of spontaneous *Drosophila* grooming.

Background

Many behaviors in animals are initiated in response to extrinsic stimuli and directed by internal neural states that control motor programs. Spontaneous innate behaviors that vary in a population and have fitness consequences are predicted to have genetic networks that contribute to variation in their expression (Mackay *et al.* 2009, McDiarmid and Rankin 2017). The interplay of underlying genetic variation and the expression of that variation through the nervous system has been incredibly difficult to parse out experimentally. The fruit fly *Drosophila melanogaster* provides a useful model system in which to address questions about the genetic underpinnings of complex behaviors, in part because of the well-characterized repertoire of behaviors (Sokolowski 2001) and genetic tools that are available to interrogate the role of genetic variants in quantitative trait expression (MacKay *et al*. 2012; Aimon and Kadow 2020). Grooming is a particularly relevant innate behavior because of its near universal widespread expression in animals and diverse purposes that contribute to an animal's fitness (Sachs 1988; Spruijt *et al.* 1992). Moreover, grooming variation in the fruit fly has been shown to be influenced by many different genes (Yanagawa *et al*. 2020) and neural networks (Seeds *et al.* 2014; Hampel *et al.* 2020). Interestingly, grooming is also a recurring exacerbated phenotype in different fly models of human neurodevelopmental disorders such as Fragile X syndrome (FXS) (Andrew *et al.* 2021). One recurring dilemma in studies that seek to study grooming and understand the neurogenetic mechanisms that contribute to variation in expression is the time-intensive process of quantifying grooming bouts. In this study, we seek to optimize the use of commercially available behavior video tracking software to increase the throughput of grooming behavior analysis.

Drosophila has long been used as a genetic model system to study human neurodevelopmental disorders (Restifo 2005; Inlow and Restifo 2004). Recent work has shown

that grooming is often aberrantly expressed in neurodevelopmental mutants, specifically those that model human Fragile X syndrome (FXS) (Andrew *et al.* 2021). FXS is a human neurodevelopmental disorder that occurs due to a loss of function mutation of the *Fragile X Mental Retardation 1* (*FMR1*) gene. It is notable in part because it was the first known example of a trinucleotide repeat disorder, which occurs due to the excessive repetition of three nucleotides CGG in or around a disease-causing gene (Garber *et al.* 2008). A complete mutation involves as many as 200 repeats of the CGG trinucleotide sequence leading to transcriptional inactivation of the Fragile X Mental Retardation protein (FMRP) (Garber *et al.* 2008). FMRP plays a key role in regulating protein interactions at neuronal synapses and abnormalities in FMRP affects synaptic plasticity. At the synapses, FMRP has been proposed to be a negative regulator of protein synthesis stimulated by group 1 metabotropic glutamate receptor (mGluR) activation (Bear *et al.* 2004). Thus, the absence of FMRP results in excessive glutamate receptor internalization in response to mGluR signaling (Nakamoto *et al.* 2007). The altered synaptic framework because of FMRP-RNA binding becomes the root cause of dendritic abnormalities, ultimately giving rise to FXS phenotypes. A key neurological feature of individuals with FXS is the presence of immature dendritic spines (reduced dendritic spine length and abnormal shapes) in regions of excitatory synapses, an abnormality that is similar to those associated with other forms of intellectual disability (Irwin *et al.* 2001; Kauffman *et al.* 2000). Nuanced understanding of the molecular pathology of intellectual disorders calls for appropriate model systems that can help us perform more extensive molecular and behavioral assays to understand intellectual disorders. The analysis of aberrant behaviors in model systems like flies is central to this approach.

The fruit fly is an ideal model organism for studying genetic neurodevelopmental disorders due to the rich repertoire of behaviors it offers, the genetic tools to identify molecular pathways associated with disease (e.g. Coll-Tané *et al.* 2019), and the low-cost maintenance, propagation, and screening available for fly labs (Qiao *et al.* 2018; Pandey *et al.* 2011). Most *Drosophila melanogaster* genes are evolutionarily conserved with humans, making fly models ideal for understanding common human diseases and behavioral disorders (Huang *et al.* 2014; Inlow and Restifo 2004). The *Drosophila melanogaster mental retardation 1* (*dfmr1*) gene is a homolog to the human *FMR1* gene and exhibits high sequence similarity with all three human FXS genes *Fmr1, FXR1*, and *FXR2* (Specchia *et al.* 2017). The dfmrp protein encoded by the *dfmr1* gene has highly conserved domains similar to human FMRP (Specchia *et al.* 2017). Fruit flies that lack the *dfmr1* gene, and hence lack dfmrp, are a genetic model for studying FXS (Dockendorff *et al.* 2002). Interestingly, they exhibit aberrant phenotypes relevant to FXS in the form of excessive grooming (Andrew *et al.* 2021). In addition, at the cellular level, *dfrm1* mutants also present defective neuronal architecture and synaptic function (Specchia *et al.* 2017; Michel *et al.* 2004). The excessive grooming exhibited by *dfmr1* mutants may be analogous to the excessive repetitive behaviors associated with FXS and other human neurodevelopmental disorders (Andrew *et al.* 2021).

In addition to helping elucidate the behavioral consequences of genetic diseases, flies also present opportunities to study the underlying neurogenetics of typical spontaneous behaviors. The *Drosophila melanogaster* reference panel (DGRP) is a publicly available resource that was developed to aid in identifying specific genetic regions and single nucleotide polymorphisms (SNPs) responsible for contributing to quantitative traits. It comprises 205 fully sequenced inbred fly lines that have been derived from a single outbred population (Mackay *et*

al. 2012; Huang *et al.* 2014). The DGRP contains a representative sample of naturally segregating genetic variation and comprises an ultra-fine-grained recombination map suitable for localization of causal genetic variants for virtually any quantitative trait (Mackay *et al.* 2012). An advanced understanding of the *Drosophila* genetic network made possible with the DGRP would complement behavioral studies showing the external manifestation of mutations in the form of excessive grooming. We utilized videos of flies from the DGRP in our current study.

Grooming is an internally programmed behavior that can be spontaneous or stimulated and is used by fruit flies to clean their body parts (Szebenyi 1969; Phillis *et al.* 1993). Studies in the past have targeted specific brain regions to gain an experimental understanding of the neural circuits that control innate cleaning programs (Seeds *et al.* 2014). For example, expression of the temperature sensitive cation channel dTRPA1, which selectively activates neurons, in different groups of brain cells caused grooming behavior in the absence of dust (Hamada *et al*., 2008; Pfeiffer *et al*., 2008; Jenett *et al*., 2012). Simultaneously, blocking cation channels in these same cells with *UAS-shibire*, which reduces neuronal activation, did not inhibit dust-induced grooming, suggesting that targeted neurons were *sufficient* but not *necessary* to drive respective cleaning movements (Seeds *et al.* 2014). Later studies have also focused on the localization of neural networks that control grooming and how intervening in the functional organization of the neural network produces changes in grooming patterns. A study by Hampel *et al.* showed that neurons are functionally connected to form a circuit that detects displacement of the antennae via mechanosensory neurons and then commands grooming through three different interneuron classes (Hampel *et al.* 2015). In all of these studies, one of the bottlenecks for studying the genetic and neural aspects responsible for spontaneous or sensory-responsive grooming was the quantification of grooming behaviors.

Grooming behavior in fruit flies has typically been scored manually. In early studies, systems were used where an observer watched a fly groom and audibly called out symbols to distinguish the various types of grooming (Szebenyi 1969). These systems were later replaced by the post-production analysis of video recordings of flies behaving either spontaneously or prompted to groom by the application of dust (Seeds *et al*. 2014; Andrew *et al.* 2021). Video annotation software is also typically employed to aid in identifying and annotating grooming bouts. The program VCode is a video annotation software package that has been used in a wide range of studies to quantify animal behaviors. It has a wide range of advantages including the support of multi-video stream, support of wide range of behaviors and events, chronological viewing of events, multiple playback modes, and consolidation of all administrative features in a single window (Hagedorn *et al.* 2008; VCode https://social.cs.uiuc.edu/projects/vcode.html). Andrew *et al.* used VCode to manually quantify fruit fly grooming in recorded videos of different neurodevelopmental mutants (2021). The downside of this manual scoring approach is that the process is time intensive and subject to scorer bias. In addition, these techniques require intensive training (Feldbauer 2020).

Automated analysis of grooming is a more efficient approach as it allows for more data collection in a shorter time interval. Some automated scoring techniques of quantifying grooming use machine learning techniques to make a computer perform a task by familiarizing it with an algorithm. A study by Qiao *et al.* developed a k Nearest Neighbor (kNN) algorithm that performs a fruit fly video image analysis of categorizing video frames into different behavioral categories after having been trained by a training dataset (2018). Every frame of an uploaded video was analyzed to see if it met the threshold value for the number of pixels to be binned into a certain behavior frame. If the threshold was not met, the background was updated, and if the

threshold was met, the background was subtracted (Qiao *et al.* 2018). The image of the fly's body was split into the core (thorax and abdomen), centroid, and periphery (legs, head, and wings) (Qiao *et al.* 2018). The change in pixels of each frame was expected to exceed or be equal to the threshold value of change of 12 because at that number the accuracy (ratio of the correct grooming levels to the output levels) of the algorithm was equal to sensitivity of the algorithm (Qiao *et al.* 2018). The algorithm provided 90 to 95% accurate results, with 5% of cases where either the grooming bout was not identified or misidentified as locomotion (Qiao *et al.* 2018). A similar approach was used by Kain *et al.* to track grooming behavior and changes in gait in *Drosophila* (2013). First, there was a set of manual scores that were used for agreement between scorers through an interrater reliability test (with only 71% interrater reliability). The machine learning scores yielded though kNN were 66% accurate. In each of these cases, very specific and expensive recording equipment was required for their analysis, and the computational expertise required to implement their detection algorithms was extensive.

This study aims to utilize a commercially available video tracking software package called EthoVision XT® (Noldus Information Technologies, Wageningen, the Netherlands) to accurately identify grooming behaviors in flies. This software is much more user-friendly than previously published automated behavioral analysis options, has a robust manufacturer support system, and is widely utilized in other animal model systems. The goal of this project it to determine the optimal settings of the software EthoVisionXT® to detect grooming behaviors. This would in turn enable an automated analysis approach to behavioral quantification that would allow for higher throughput and increased repeatability between experiments.

Materials and Methods

Video recordings and fly behavior data collection

Previous lab members Courtney Hannum and Zachary High created a video collection comprising 157 10-minute-long videos that captured spontaneous behavior in >750 fruit flies from 34 different DGRP lines. These videos were collected for a study that used manual scoring techniques (described for the current study below) to examine grooming (Hannum 2017). They defined grooming as any stroking of the head, antennae, proboscis, limbs, wings, or thorax using one or more limbs (Hannum 2017). Fly behavior was recorded 20-24 hours after eclosion (when a mature fruit fly emerges from its pupal case) to make sure that all animals were at comparable developmental ages. The animals were brought to the behavior room at least an hour before the recording time to allow them to acclimate to room conditions, which were a temperature of 25℃ and 70% humidity. For each video, six un-anaesthetized flies were aspirated in individual wells of a sterile polystyrene flat-bottomed 96 well plate. The plates were pre-filled with 200 µL 1.5% of agar to maintain a consistent substrate, provide humidity, and restrict fly movement vertically in the recording arena (Hannum 2017). Videos were recorded with a Canon Vixia HF R72 camera. The camera was placed overhead of the wells and all recordings were performed in the dark under the light of a circular microscope LED desk light (Hannum 2017). Before every recording session, the name of the recorder, DGRP line and sex of the fly were noted and during the recording session patterns of behavior, such as excessive grooming, walking, standing were noted. At the end of the recordings, the flies were put in individual 0.5 mL microcentrifuge tubes and placed in a -20℃ freezer for subsequent genotyping. All videos in the current study were recorded in this manner during the summer of 2016.

Choosing representative data set for manual and automated scoring

The scoring results from the previous lab members were stored in an excel spreadsheet comprising the video number, fly line genotype, well in which the fly was aspirated, corresponding grooming bout length, grooming index, and number of grooming bouts. A grooming bout is defined as the time between the start and stop of an interval of uninterrupted grooming. Thus, each video can be broken down into numerous bouts of walking, standing, and grooming. Grooming index (GI) is the percentage of the total time the fly spends grooming (sum of all grooming bout lengths divided by length of video multiplied by hundred). For our project, we used a random number generator (www.random.org) to assign random numbers to each of the scored flies from this previous study. We chose the flies with the randomly-assigned numbers 1-50, out of the total set of >700 previously scored animals, as our working set moving forward. This approach was employed to allow us to choose animals exhibiting a diverse array of grooming patterns and reduce any genotype bias to our analysis. We then entered all the information (grooming indices, mean length of bout, and number of grooming bouts) for this set of animals into a new excel spreadsheet that was then filled in with the corresponding information from the current user.

Manual Scoring with VCode and Interrater Reliability Testing

From the working set of 50 flies, each corresponding video was loaded into the VCode video annotation software (Hagedorn *et al.* 2008). Each of the 6 wells in the video was assigned a specific key to indicate when grooming bouts occur. To score and quantify grooming, the user observed the video and pressed the designated key once the fly started grooming and pressed it again when the fly stops grooming, thereby demarcating a grooming bout in the video. This was repeated for every grooming bout in a given video for the extent of the 10-minute recording. The

space bar was used to pause the video, and the left and right arrows could be used to navigate frame-by-frame to accurately mark bout start- and stop-points. Once all 50 videos were scored, we used custom in-house Perl scripts to extract grooming data from the VCode event log outputs. The extracted data comprised the durations of grooming bouts, number of bouts, and GI for each fly. After all the data was put into the excel spreadsheet, we focused on figure construction and statistical analysis. A custom MATLAB script read the text-based VCode event log outputs to generate the ethogram figures that graphically illustrate the time and duration of grooming bouts during the 10-minute observation window. We compared the ethograms of the current user to the previous users to note differences in observed grooming bout length and number of bouts, which also corresponded to differences in GI. This was our interrater reliability (IRR) assessment to ascertain significant and persistent differences in the identification of grooming bouts. We performed a Wilcoxon Rank Sum test between the current and previous users' data to generate p values to determine if there are statistical differences (i.e. $p < 0.05$) between the scorers in each metric. After comparing the results from the IRR test, we chose a working subset of ten animals to initiate our automated scoring pipeline. We needed to pick a reasonable number of animals and ten seemed like a good number for that. We chose the animals with good congruence between manual scores of all scorers in order to have a confident benchmark against which we could compare subsequent automated methods.

Automated Scoring with EthoVision XT®

Background calibration, arena, and detection settings

EthoVisionXT® (hereafter referred to simply as EthoVision) is a commercially available behavior tracking software platform (https://www.noldus.com/ethovision-xt – Noldus Information Technologies, Wegeningen, the Netherlands). Although it has previously been

utilized to observe fly movement for different purposes (i.e. Kaur *et al.* 2015), we sought to implement it for automated detection of grooming-specific behaviors, building off of the work of previous research students (McLaughlin 2018; Feldbauer 2020). EthoVision works by first defining a behavioral arena in a video file in which an animal is behaving, then employing user-defined pixel-value based identification of the subject to identify the location, size, and frame-by-frame movement of the animal within the arena. The user is responsible for defining the location, size, and shape of the arena, establishing the detection threshold settings for identifying the subject in each video, and providing settings to determine the different behavior metrics of interest. Based on previous work in the lab (Feldbauer 2020), we focused on varying two different detection variables, activity and movement (described below), in the current analysis to attempt to optimize the identification of grooming.

We started with the working subset of ten flies with the most congruent interrater reliability scores. The first step in EthoVision is to prompt the user for a brief description of the experiment. The user then calibrated the scale of the arenas by defining the distance between three wells of a 96-well plate at 2.6 cm (**Figure 1**). Background calibration also allows EthoVision to differentiate the background from the animal in the foreground and focus on the arenas containing the flies to be scored. An arena was created by drawing a circle of the exact same size over the well containing the fly of interest. The video trial settings were then checked for validation and saved. A common trial control setting was set at 600 s after arena settings had been completed for all the animals to limit the analysis to 10 minutes. This was done in correspondence to the manually scored videos that were ten minutes long. The final step for making the videos ready for data acquisition was to make sure that the animals would be detected properly.

Under the automated setup option of the detection settings, we delineated the fly body from legs and wings by drawing a box around the body of the fly (**Figure 2A** and **B**). To fine tune the settings, under the advanced setting option, we utilized the contour erosion and dilation to make sure that no legs or wings were being recognized as the body. These settings helped distinguish the pixels that define the body (dark pixels with low grayscale values, minimum 0) from the arena's background (light pixels with high grayscale values, maximum 255) and the animal's appendages (Feldbauer 2020). The center of the fly's body was calculated as the pixel occupying the center of mass of the detected body outline. In addition to these parameters, we also altered the **initial acquisition activity setting** (focusing on a range of 15 to 17), which determined the pixel value change required to categorize a pixel as changing state from one frame to the next (See **Table 2**). The purpose of adjusting the initial activity setting was to clearly identify the movement of a fly's legs in relation to the background while simultaneously not identifying pixels that are not associated with the fly as changing (i.e. reducing background noise) (**Figure 2C**). This logic should identify the appendage movements of a non-walking fly.

Data acquisition and data profile creation

After specifying the required settings and putting all the videos and their corresponding arena and detection settings in a trial list, the entire video subset (n=10 flies) was ready to be automatically scored by EthoVision. An advantageous feature of EthoVision is that the settings can be modified as to allow for one video to get scored after the other without requiring the user to intervene in the process, thus saving time and energy. Behavior acquisition consists of the software determining the location of the center point of the animal, the outline of the animal, and the number of pixels in the arena that change values from frame-to-frame, all based on the initial detection settings established above (**Figure 2**). Post data-acquisition steps then allow different

data profiles with varying **activity** and **movement threshold** parameters (See **Table 2**). These were the metrics we varied to see what parameter combinations yield the best automated scores. A previous student established a behavior-binning logic that includes two parameters, **movement** and **activity**, to classify video frames as walking, standing, and grooming (**Figure 3**; Feldbauer 2020). Both parameters were assigned a certain threshold value for the purpose of behavior classification. *Movement* is defined as the distance change in the center point of the animal from frame to frame. If the change in center point of the animal (*movement*) exceeded a defined threshold (our **movement threshold**), then that the animal would be classified as "moving" during that frame of the video. Activity is defined as the percentage of pixels in the defined arena that change values beyond a the initial activity acquisition threshold between two adjacent frames. The activity setting that we altered was the percentage of these pixels in the arena. If the threshold value for movement was not exceeded (indicating the fly is not walking), then EthoVision would check if the percentage change of number of pixels in the arena are exceeding the threshold value set for the **activity threshold**. If the change in percentage of pixels exceeds the activity threshold, we considered this a grooming bout under the logic that if the animal is not moving (i.e. sub movement threshold value) but there is a large percentage of pixels changing value in the arena (i.e. above the activity threshold value), then the animal's appendages are changing position, and it is therefore likely engaged in grooming. If the animal is not moving and the activity is also below our activity threshold, then the fly was considered to be standing. The main focus of this project was to use different **parameter sets**, specific combinations of movement and activity thresholds, to determine what combination yields the most accurate automated scores as compared to our manually-scored benchmark results.

For the working subset of ten videos, we generated data at three initial activity acquisition settings of 15, 16, and 17. At each of these acquisition settings for activity, we acquired data at movement thresholds of 0.05 and 0.07 cm/s with activity thresholds ranging from 0.10 to 0.35 (i.e. 10-35%) in increments of 0.05. Previous experiments helped us determine that going above 0.35 and below 0.05 activity threshold settings yielded scores with substantial statistical significances from the manual scores. The results from the working subset helped us identify a focused set of software settings to focus on for further optimization: initial activity acquisition setting of pixel-value change equal to 16 with movement threshold 0.05 cm/s and activity threshold 0.10 and 0.09 (a range of 9% to 10% change in pixels); movement threshold of 0.07 cm/s and activity threshold of 0.15 and 0.16. These settings were then applied to the entire set of 50 animals.

Python Pipeline and Statistical Analysis

For every parameter set, EthoVision exported a Microsoft Excel (.xlsx) file containing the cumulative durations of grooming, walking, and standing in seconds and percentage of total time the fly spends performing a certain behavior (indices). These files were put through a data analysis pipeline containing custom Python scripts (created by Mikayla Feldbauer) for extracting relevant population statistics. One of the preliminary steps in the pipeline was to convert the Microsoft Excel files into comma separated value files needed for extraction using Python (Feldbauer 2020). A key feature of Python that eases data analysis is the use of classes. A class in Python allows a segment of code to be reused for same operations but different data every time it is called. For each parameter set data, we called the EV_Analysis_SOHO class to perform the same calculations, such as generating the differences between manual and automated scores,

creating box plots of scores for the animals, implementing Kruskal Wallis and Mann Whittney U test for significance, and performing Ordinary Least Product (OLP) regression analysis.

Two different methods of measurement can never be perfect because of inconsistencies or biases in the method of measurement or on the part of those utilizing the method of measuring (Ludbrook 2010). We had two main reasons to perform a statistical analysis in our study: to calibrate one method against the other (i.e. automated vs. manual scores) and to detect bias in our automated scoring regimen (Ludbrook 2010). Through the Python pipeline, we performed an OLP regression analysis to detect any bias between the two methods of manual and automated scoring. An OLP regression plot comprises a regression line and an identity line. The y axis consisted of all the manual scores and the x axis those of the automated scores. The identity line is our theoretical target, where every y value is equal to the x value. This would indicate that every EthoVision generated automated score is equal to the corresponding manual score for each animal. A regression line is the best fit of data points and determines how close the automated scores are to the manual scores. OLP regression does not assume that one score is necessarily "correct" because both methods of scoring (i.e. manual and automated) are expected to result in some errors. The equation of the OLP regression line as determined in the Python pipeline provides values for the y intercept (indicated as ***a***) and the slope (indicated as ***b***). Bias can occur in two types: 1) fixed bias, where one method gives values that are consistently higher or lower than the other method, and 2) proportional bias, where one method gives values that are higher or lower than another method by an amount that is proportional to the level of the measured variable (Ludbrook 1997). Bias from an OLP regression plot can be understood with the help of confidence interval of the regression line. A confidence interval allows for an estimate of certainty around the parameter of interest (Perry *et al.* 2017). The a_CI95 and b_CI95 values

provided by the OLP function we employed are the 95% confidence intervals for *a* (y intercept) and *b* (slope), respectively. For our graphs, the *a* was the intercept and the *b* was the slope. A fixed or systemic bias is present if a_CI95 does not include the value 0, indicating that the regression line does not run through the origin and suggests a value higher (or lower) across the whole range of measurement. A proportional bias, i.e. when one method gives values that diverge progressively from those of the other, is present if the 95% confidence interval of the slope (b_CI95) does not include the value 1.

Correlation is another statistical technique that can show how related two variables are to each other (Giavarina 2015). Bland and Altman developed such a technique that would help in replacing an old method with a more advanced method (in our case, replacing manual with automated scoring) by looking at the scatterplot of the values within a population where the y axis is the difference in scores between the two methods for each fly and the y axis is the mean of the two measurements (Bland and Altman 1986). A Bland Altman plot appears like a normal distribution plot turned on its side. The central line is the mean and then we have intervals that are within 1 standard deviation followed by 2 standard deviations. Bland and Altman recommend that 95% of the data points lie within ±2 standard deviations of the mean difference, suggesting a 95% confidence interval (Giavarina 2015). The in-house Python scripts called specific classes that generated Bland Altman plots for each of the parameter set results.

Good statistical practices demand reporting of some measure of variability or reliability for important statistical estimates (Boos and Stefanski 2011). Therefore, our final analysis included generating p values using Mann Whittney U and Kruskal Wallis tests. Mann Whittney U, also known as Wilcoxon Rank Sum test, is a non-parametric test that is performed on a large sample where assumptions of normal distribution are questionable (Rosner and Grove 1999). In

the evaluation of this pairwise test, each data point from one sample is compared to that of the second sample (Feldbauer 2020). The Kruskal Wallis test is also another non-parametric test that aims to check if two samples have originated from the same population (Guo *et al.* 2013). The biggest difference between the two tests is that Kruskal-Wallis test can accommodate more than two data groups whereas Mann-Whitney can accommodate exactly two data groups (Hazra and Gogtay 2016). We carefully analyzed the results generated by each of the statistical techniques and drew conclusions about optimal parameter settings.

Results

IRR demonstrates the difficulty of accurate grooming identification

Our interrater reliability test provided us with manual scores for GI that were ready to be compared to automated EthoVision-generated scores. We first sought to determine if scoring for grooming behavior was consistent between previous members of the lab and the current author. The data for GI, number of grooming bouts, and average (mean) bout length were first visualized with box-and-whisker plots to view the aggregate population of scores for two different sets of scorers for this subset of 50 animals (**Figure 4**; "CZ" represents the scores of Courtney Hannum and Zachary High (Hannum 2017), "SM" is the current author). The white lines in the middle of the boxes represent the median and the boxes represent the interquartile range (25^{th} to 75^{th} percentile), whereas the whiskers on the top and bottom represent 90^{th} percentile and 10^{th} percentile, respectively. We observed no statistically significant differences in the current user's grooming indices and number of bouts when compared to previous student scorers (**Figure 4A and 4B**; Wilcoxon Rank Sum test $p > 0.05$ for both GI and number of grooming bouts, $n = 50$). However, there was a difference in the mean bout length (**Figure 4C**, Wilcoxon Rank Sum test p

< 0.05, n = 50). In the OLP regression plot, we noted how the regression and identity lines for GI were in close proximity to each other, and the 95% confidence intervals for the intercept and slope support the similarities between scorers for grooming index, but with a slight proportional bias noted (**Figure 4D**; a_CI95 for intercept = [-0.1416 0.3840] and includes 0; b_CI95 = [1.0441 1.1355] and does not include 1). We used the regression plot (**Figure 4D**) to select 10 animals that had manual scores of the current user most congruent to the manual scores of the previous user to create a smaller working subset on which to perform the first screening steps for automated analysis.

We sought to determine the factors driving the difference in bout length between scorers by comparing ethograms of animals with disparate scores (**Figure 5).** An ethogram shows the timing and duration of individual grooming bouts over a ten-minute video interval in a graphical way. In each of the ethograms, the start and stop point of grooming for all the bouts remained very similar for all the videos, suggesting constant maintenance of the grooming indices. The number of bouts indicated by the number of black lines remained statistically comparable as well. The only parameter that showed a difference, albeit minor, was the mean bout length. For example, in ethogram **C** (**Figure 5C**), the last bout in SM was broken into multiple smaller bouts of grooming whereas the previous user CZ had scored the bout as one constant bout. Similar bout-breaking, due to a cumulative effect, resulted in an increase in the differences between users' mean bout length. We looked further into bouts that differed between scorers (as indicated by red asterisks in figure 5) and determined that for many of these bouts there were indeed ambiguous movements or brief grooming breaks that resulted in the differences between scorers. These in-depth observations illustrate the difficulty in agreeing upon all grooming bouts, even between well trained scorers. Ultimately, however, we determined that the similarity of the

summary metric of GI between scorers, as supported statistically (**Figure 4A** and **4D**) and graphically (**Figure 5**), supports the validity of our manual scoring approach to act as a benchmark for verifying the automated analysis we are attempting to optimize. We focused exclusively on GI moving forward for automated analysis.

Initial activity setting of 16 yielded the best results in initial screens

We began our EthoVision screen by setting up our pipeline with the 10 animals chosen above that had unambiguous grooming bouts as determined by detailed comparison of ethograms from several scorers. The purpose of this initial screen was to learn the nuances of the analytical pipeline and quickly determine ranges of parameter settings that would be analyzed in further detail with the full set of 50 animals. In a previous Honors project, Mikayla Feldbauer (2020) determined that an initial acquisition activity setting range of 15 to 20 should be used as the appropriate detection settings. This led us to question: is there a best detection setting out of these 6 that could generate more accurate results? We therefore started our experimental setup with inidial acquisition activity settings of 15, 16, and 17. All three initial acquisition activity settings yielded the statistically insignificant differences between manual and automated scores at a movement threshold of 0.05 cm/s and an activity threshold of 10% (0.10) (Figures **6A–C**). However, for a movement threshold of 0.07 cm/s, the three initial acquisition activity settings had different activity thresholds that generated automated scores comparable to manual scores (**Figure 6D–F**). At an initial acquisition activity setting of 15, an activity threshold of 10% yielded the least statistically significant automated scores (**Figure 6A** and **D**). An activity threshold of 15%(0.15) produced statistically insignificant differences between manual and automated scores for initial acquision activity setting of 16 and threshold of 0.07 cm/s (**Figure 6E**). Finally, an activity threshold of 20%(0.20) at an activity setting of 17 yielded the least

statistically significant differences between manual and automated scores for a movement threshold of 0.07 cm/s (**Figure 6F**). In all cases, the last set of activity thresholds of 30 and 35% (as indicated by the red asterisks) yielded statistically significant differences between manual and automated scores (Wilcoxon Rank Sum test $p < 0.05$, $n = 10$). We further observed that the differences yielded by the initial acquisition activity setting of 17 had outliers indicating larger differences between automated and manual scores (**Figure 6F**). We therefore narrow our choices between initial acquisition activity settings of 15 and 16. To simplify the next steps of our screen, we decided to focus on an initial acquisition activity setting 16 because modifying this setting requires adjusting the settings for each video in a set. Initial acquition activity settings are one of the more arduous variables to change in this analytical pipeline because it requires setting up different acquisitions of data in EthoVision for each video. We reasoned that choosing one initial activity setting would enable us to do a more complete sensitivity analysis of many different movement and activity thresholds on both our subset ($n = 10$ animals) and complete set ($n = 50$ animals). In addition, some of our manual versus EthoVision difference boxplots supplemented our reasoning by depicting that the differences between the EthoVision and manual scores were the least for an activity setting of 16.

Sensitivity analysis yielded successful results in a preliminary subset of animals

After applying the initial acquisition activity setting of 16 to the initial subset of ten animals, we first identified two specific parameter set combinations that were yielding differences between manual and automated scores (manual – automated score = 0) that were tending toward zero at two different movement thresholds of 0.05 cm/s (**Figure 7A**) and 0.07 cm/s (**Figure 7D**). Movement threshold of 0.05 cm/s and activity threshold of 0.10 (i.e. 10%) (**Figure 7A** and **B**) and movement threshold of 0.07 cm/s and activity threshold 0.15 (i.e. 15%) (**Figure 7D** and **E**) showed the best results for our initial ranges of activity settings. Before we

proceeded with applying these settings to the entire set of 50 animals, we wanted to confirm the settings that we had identified. Therefore, for the movement threshold of 0.05 cm/s, we focused on the activity threshold range of 0.05, 0.06 0.07 0.08, and 0.09 (5% to 9%); for the movement threshold of 0.07 cm/s, we focused on an activity threshold range of 0.13(13%) to 0.17(17%) in 1% or 0.1 increments (**Figure 7C** and **F**). We observed that under the focused parameter sets, for a movement threshold of 0.05 cm/s, an activity threshold of 0.09 (i.e. 9%) yielded indices that were the most similar to manual scores (**Figure 7C**, Wilcoxon Rank Sum test $p = 0.484$, $n = 10$). Similarly, for a movement threshold of 0.07 cm/s, an activity threshold of 0.16 (i.e. 16%) yielded automatic indices most similar to automated scores (**Figure 7F**, Wilcoxon Rank Sum test $p = 0.484$, $n = 10$). These activity threshold differences that were off by just 0.1 when the activity threshold ranges were focused on, indicated that there could be potential scaling errors in the graphs generated by Python that could be responsible for the differing parameter settings. We chose the four combinations as shown in **Table 1** that yielded least statistically significant automated scores (i.e. highest p value under the Wilcoxon Rank Sum test) and proceeded with these values for the full set of 50 animals. We excluded the other activity thresholds because the automated scores were significantly different from manual scores (indicated by red asterisks in **Figure 7**).

Parameter setting analysis converges on manual scores for a larger population

When the software setting combinations from **Table 1** were applied to the entire set of 50 animals, movement threshold of 0.05 cm/s and activity threshold of 10% yielded the least differences between manual and automated scores in both percent difference (**Figure 8A**) and population GI measures (**Figure 8B**). The box plots of grooming indices show these results clearly because the median of the manual scores was closest to the automated scores for the box plot at a movement threshold of 0.05 cm/s and activity threshold of 10% (0.10) and the

interquartile range boxes were most overlapping (**Figure 8B**). A movement threshold of 0.05 cm/s and an activity threshold of 10% (0.10) yielded the least statistically different automated scores (Kruskal Wallace p value = 0.276; **Figure 8**). The movement threshold 0.07 cm/s along with activity thresholds of 15% (0.15) and 16% (0.16) showed significant underscoring of automated scoring when applied to the entire 50-animal data set. The distribution of percent differences in manual and automated scoring obtained at a movement threshold of 0.07 cm/s was also more spread out than the data obtained at 0.05 cm/s (**figure 8A**). The results were confirmed by OLP regression and Bland Altman plots (see **Figure 11**).

The OLP regression analysis for the movement threshold and activity threshold combination of 0.05 cm/s and 10% (0.10) (parameter setting 2) contained closely aligned regression and identity lines (**Figure 10A**). A small deviation from the identity line (x = y) indicated potential biases in EthoVision scoring. The absence of 0 in a_CI95 indicated the presence of potential proportional bias (see Discussion). This was also confirmed by a positive correlation (graph showing a positive trend of points; increase if x with increase in y) of data points on the corresponding Bland Altman plot (F**igure 10B**). Because the OLP regression line was consistently above the identity line, we can infer that EthoVision constantly overcalls grooming behavior for this specific parameter set, which was suggested by the population boxplots of percent difference (**Figure 8A**) and grooming index (**Figure 8B**). The positive correlation of the data points on the Bland Altman confirms the proportional bias (**Figure 10B**).

Potential conflict arises in different forms of comparison for the best parameter settings

After narrowing down the parameter settings, we decided to have one final parameter setup that would span over the entire range of activity thresholds from 8% (0.08) to 16% (0.16) at the movement threshold of 0.05 cm/s so that we could observe the percent change in differences from negative to positive and thus find the best fit. We observed that at a movement

threshold of 0.05 cm/s an activity threshold of 10% (0.10) yields the boxplots that illustrated the least percent difference between manual and automated scores (**Figure 9A**) when the automated GIs were directly compared to the manual scores for each animal. Interestingly, an activity threshold of 14% (0.14) yielded automated GI scores that were least statistically significant from manual scores at the population level (**Figure 9B**). The setting of 14% (0.14) also yielded the highest Kruskal Wallace p value of 0.945, indicating that the GIs of the 50 animals were quite similar at the population level. Moreover, the absence of 0 in a_CI95 indicated a proportional bias and was also confirmed by a positive correlation of data points in the Bland Altman Plots (**Figure 10C** and **D**). In this set of parameter settings, the activity threshold of 8% (0.08) yielded the smallest p value ($p = 0.015$), and therefore worst settings. We therefore decided to generate OLP regression and Bland Altman plots to confirm if the differences are indeed significant. The absence of 0 in a_CI95 indicated the presence of proportional bias as was also confirmed by the positive correlation of the Bland Altman plot (**Figure 10 D**), thus confirming the poor fit of this parameter setting.

There is, therefore, a conflict in how our data are expressed that shows that the most accurate settings on an animal-by-animal basis, as determined by the distributions of GI percent differences between manual and automated scoring, is different than the parameter values that provide the least difference at a population level. That is, with a movement threshold of 0.05 cm/s the parameters of 14% (0.14) for activity threshold deliver the best fit of the data when considered as a population (**Figure 9B** and **10C, D**), but the activity threshold of 10% (0.10) was the best for accurately determining the GI for each individual animal (i.e. reducing the % difference) (**Figure 9A, 8B** and **10C, D**).

Discussion

Manual scores as a benchmark for automated scores

For our project, we utilized manual scores as a benchmark to compare the semi-automated scoring results of EthoVision. There were differences in the manual scores between the current and the previous users due to a difference in perception of grooming bouts. Previous users of VCode were scoring fve to six flies in one video whereas the current user was focused on one single fly under one video. That allowed the current user to pickup minute details of when a fly starts and stops grooming. The method of scoring or defining a grooming bout was likely the underlying reason as to why the current user showed a marked increase in the number of bouts and overall shorter bout lengths. However, it is important to note that no method is better than the other or more "correct." There are tradeoffs and biases associated with both methods, and the IRR was a way to gauge if there were any stark differences between the current and previous users. We were not under the assumption that the manual scores were "correct." We were simply using them as a set up to see if we can generate matching semi-automated scores with commercially-available software. There is a potential that EthoVision scores are more consistent than manual scores and thus extensive benchmarking and statistical comparisons can ultimately lead us to an understanding of the subtle differences between these methods.

Successful parameter settings and benefits of the automated system

Our project has identified a set of software settings that yield accurate automated scores that are not statistically significant from manual scores at both the individual animal level and the population level. At a population level, a movement threshold of 0.05 cm/s and activity threshold of 14% (0.14) yielded the least statistically significant differences between automated and manual scores. At the individual animal level, a movement threshold of 0.05 cm/s and an activity threshold of 10% (0.10) yielded grooming indices that were most similar to manual

scores. In each of these cases, the p values for both the Mann Whitney and Kruskal Wallis statistical tests were greater than 0.05, thus confirming that there were no statistically significant differences between automated and manual scores. The p value for an activity threshold of 14% (0.14) along with a movement threshold of 0.05 cm/s was the highest ($p = 0.945$, $n = 50$). Our results affirm that population values can be sporadic. Throughout the project, there were no instances of systemic bias for the parameter sets that we chose, thus indicating that EthoVision is accurately capable of determining grooming behavior bouts. The other main benefit of our approach is that there is no extensive computing and set up of complex algorithms that were used in methods such as the kNN algorithm (Qiao *et al.* 2018). That saves a lot of time on familiarizing a computer system with the working of the algorithm or a training dataset like the kNN from Qiao *et al.* (2018). For example, it took a well trained user an hour for setting up the background and detection settings for 10 animals. EthoVision ran each of the videos for ~10 minutes and then it took the user another 30 minutes to set up data profiles and export data. Once a user manual has been created for this project, a novice user can utilize EthoVision to automatically score fruit fly videos for grooming bouts.

Inconsistencies with the software and some programming loopholes

When compared to the manual scores, throughout the project there were multiple instances of proportional bias which means that EthoVision tends to constantly overscore or underscore grooming bouts with different parameter settings (the absence of 0 in the a_CI95, where a is the intercept in the equation of the line of the OLP regression graph). That could potentially be due to the initial software settings linked to the binning logic causing certain behavior frames to be misclassified as grooming. This is one downside of automated scoring. The initial software settings are difficult to maneuver and can vary from one user to another and

potentiall from one video to another leading to downstream bias and error. However, once a thorough manual/user-guide is made, these biases could be reduced. Owing to the initial settings, certain behavior frames such as an instance when a fly is exhibiting a lot of body movement could be potentially categorized as grooming due an increased change in pixels. This can exhibit a cumulative effect because if too many non-grooming movements with sufficient change is pixels are categorized as grooming, then that can lead to overscoring and thus proportional bias.

We have also come to the realization that a training set plays a crucial role in the identification of optimal software settings. With our training set of ten specific animals we had narrowed down the activity threshold to a range of 10% to 15% along with a movement threshold of 0.05 cm/s. It would be of interest to check to see what would happen if a larger training set of maybe 15 or 20 animals would be used for initial assessment of semi-automated scoring.

An important observation that we made from **Figure 7** was that when an analysis assay was run from 10% (0.10) to 35% (0.35) over 0.5 increments, at a combination of movement threshold 0.05 cm/s and activity threshold 10% (0.10) EthoVision yielded automated scores that were least statistically significant from manual scores. When a similar analysis was run at a focused and short range of activity thresholds, the setting that yields the least differences from manual scores, differed. A similar observation was made for a movement threshold of 0.07 cm/s and an activity threshold of 15% (0.15). That lead us to question if the graphs being generated by Python were actually scaled properly. If the scales of the graphs differed, then we would have a wrong parameter setting showing us the least differences between EthoVision and manual scores. This is something we seek to address in future projects.

Conclusion and future directions

Our project confirms that EthoVision can indeed be utilized for automatically scoring fruit fly videos for bouts of grooming, walking, and standing. An activity setting of 16, a movement threshold of 0.05 cm/s, and activity thresholds of 10% (0.10) and 14% (0.14) generate a range of automated scores that have been shown to be accurate indices of grooming. This indicates that EthoVision allows for the usage of multiple different parameter sets for the accurate scoring of grooming bouts.

In this project, we have identified that an initial activity setting of 16 and movement threshold of 0.05 cm/s yields results that are accurate over a narrow range of activity settings. However, the project still needs some fine tuning to identify which activity thresholds ranging from 10% (0.1) to 14% (0.14) would perform the best binning of video frames, thus yielding the best set of automated scores that align with manual scores throughout all statistical analysis techniques. This means that for that one specific parameter set, the particular box plot of differences between manual and automated scores would be close to equal to zero, the p values for Mann Whitney and Kruskal Wallis test would be the highest, and the regression and identity lines on the OLP regression plot would be in very close proximity or potentially overlap each other, with no indication of proportional or systemic bias. Some new statistical analysis technique such as a paired t test could be used to compare the results yielded by the two settings. The existing inhouse Python scripts could be modified to check what parameter set of movement and activity threshold is consistently yielding the least statistically significant differences between manual and automated scores. Moreover, future students should focus on graphically representing EthoVision-scored bouts in an ethogram to compare to manual scores in order to

identify regions of conflict comparable to the interrater analysis that we performed in this study (as in **Figure 5**).

Once the optimal software settings have been confirmed, videos of *Dmfr1* mutants displaying excessive grooming can be uploaded on the software in contrast to wild type flies from DGRP to check to see if the software is able to identify all the grooming bouts. In preparation for that, the wild type and excessive fly grooming should be first recorded and then manually scored for interrater reliability. The users checking for interrater reliability should be on the same page regarding the definition of grooming. The software has a lot of potential of being used for a variety of behavioral assays and experiments should continue to exploit automated analysis in the quest of identifying the optimal parameter settings.

Acknowledgements

These four years of undergraduate studies have been the most amazing years of growth that I cherish. I am sincerely grateful to the Biology department for all the support and kindness. I would like to thank Dr. and Mrs. Haberberger for their generous Fellowship and my research advisor, Dr. Andrew, for constantly being there for me and guiding me throughout the research process. I would also like to thank Drs. Morrison, Bendorf, and Ramsey for their valuable feedback as part of my Honors committee. A big thank you to Mikayla Feldbauer for her constant support, enthusiasm, and help over the course of the research process. I am thankful to my International Student Advisor, Mrs. Marleni Feinstein, for her invaluable guidance throughout these years. I am incredibly grateful for my parents Kaoushik and Rita Mukherjee for letting me achieve my dreams and constantly loving me through all the periods of pain and exhaustion. Last but not least, I would like to offer a token of appreciation to my best friend, Shruti Bhunjun, for holding my hand throughout college and my dear friend Soojay Jhugaroo for making me realize how amazing I am.

Figures and Legends

Movement Threshold cm/s	Activity Threshold
0.05	0.09
0.05	0.10
0.07	0.15
0.07	0.16

Table 1: Combinations of movement and activity threshold that were used for subsequent analysis after an activity setting of 16 was determined to be utilized for subsequent analysis.

Parameters	Unit
Initial acquisition activity setting	Absolute change in frame-by-frame pixel values
Movement threshold	cm/s
Activity threshold	Percentage of arena area

Table 2: The EthoVision parameters and their corresponding units that were used to set up a binning logic for automated categorization of spontaneous behaviors in the *Drosophila melanogaster*.

Figure 1: Preliminary steps in automated scoring of videos using EthoVision included background calibration and creation of arenas. Under arena and background calibration settings, we calibrated the scale of the background at 2.6 cm across three wells and drew circular arenas around the fly of interest for the software to focus on.

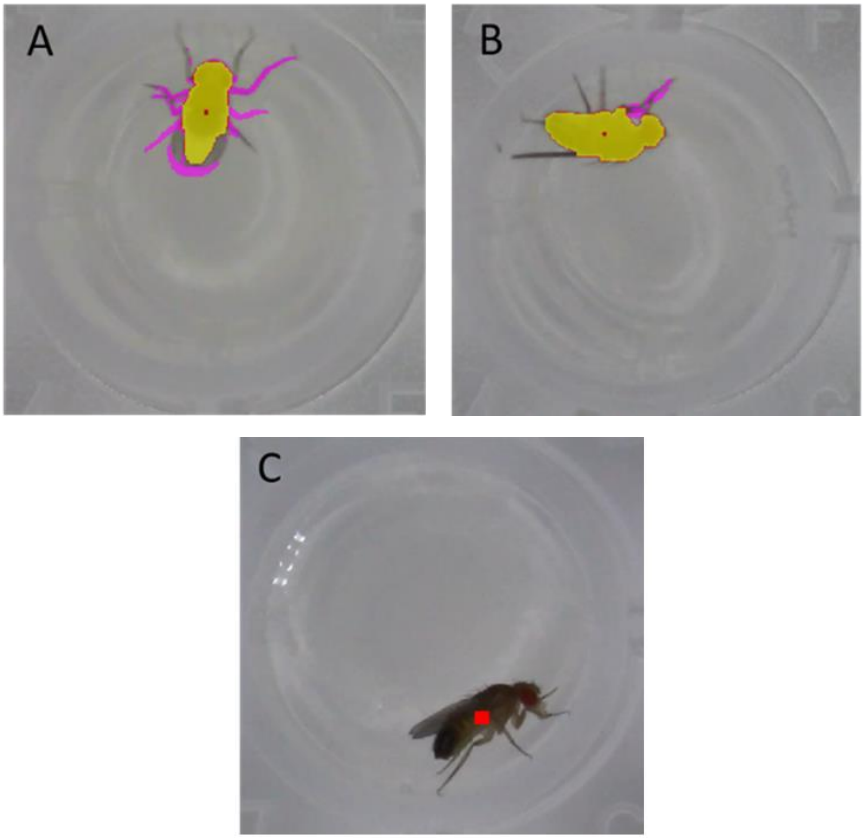


Figure 2: We used EthoVision detection settings to demarcate the fly body from the legs and wings so that the parameters of interest accurately bin the changes in center point and the percentage of pixels (A) and **(B)** show the detection settings that help in delineating the fly body from the wings and legs so that the activity threshold parameter accounts for the percentage change in pixels. **(C)** Shows the determination of the center point of the fly body movement threshold accounting for drastic changes in the center point of the fly.

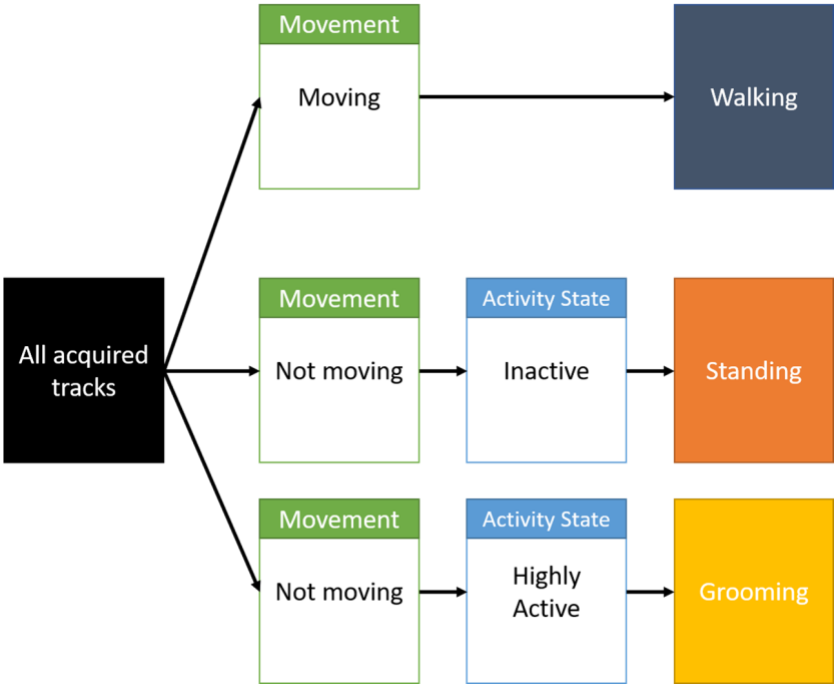


Figure 3: Logic tree used by EthoVision to bin video frames into different behavior categories (Feldbauer 2020). As inferred from the logic tree, if a fly is moving, then it is walking. If a fly is not moving but highly active, then it is grooming, and if a fly is not moving but inactive, then it is standing. The movement thresholds of 0.05 and 0.07 cm/s were determined to work well in identifying moving from non-moving. The activity state is determined by the activity thresholds that we varied in our experiments.

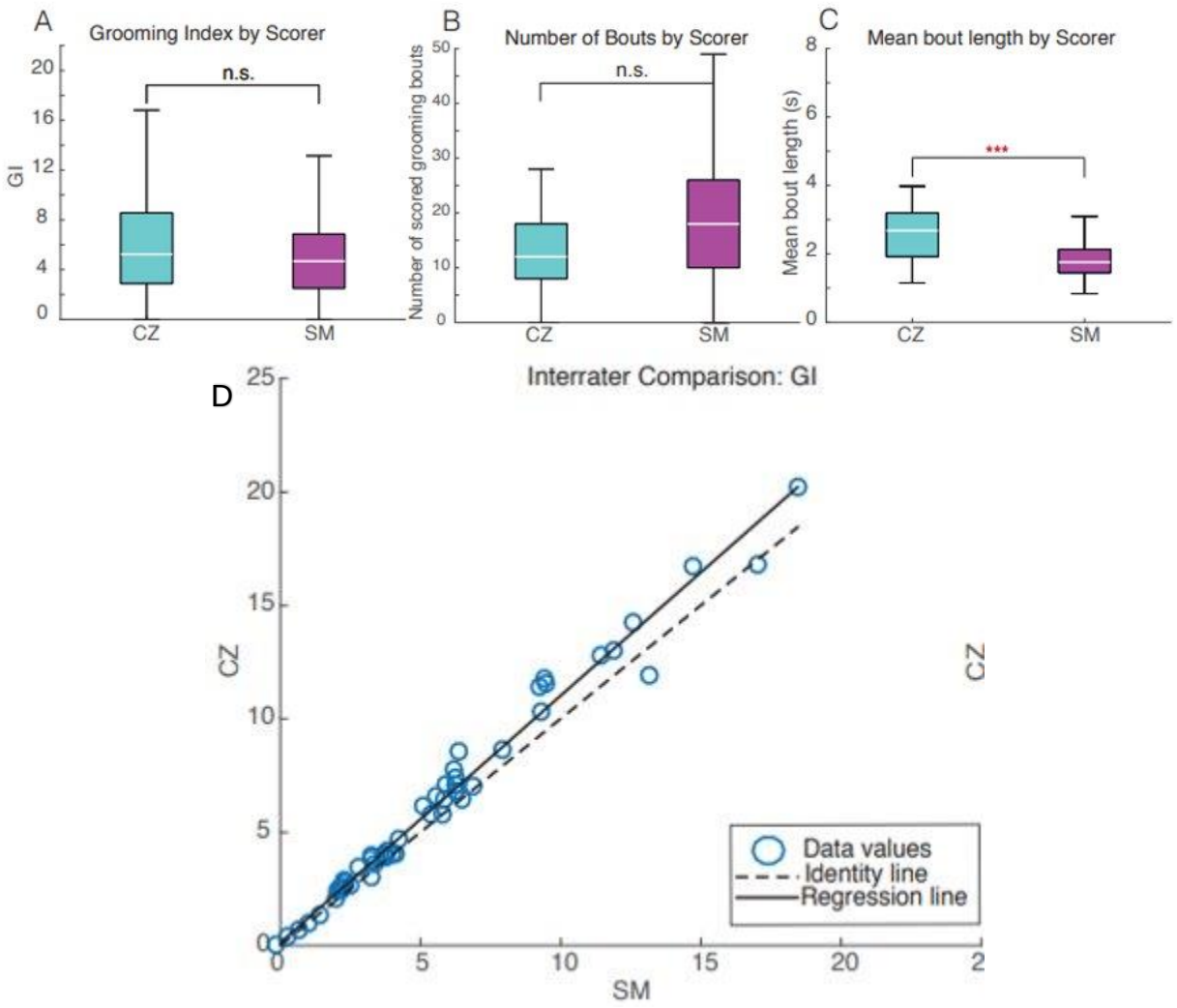


Figure 4: Interrater reliability test statistics included OLP regression, Wilcoxon Rank Sum test, and box plots of grooming indices, number of bouts, and mean bout length (n = 50). (A) There were no statistically significant differences between variations in grooming indices between the current and previous lab users of VCode. **(B)** Current user SM showed a marked increase in scoring the number of grooming bouts in each video **(C)** There was a large statistically significant difference in between the previous and current users in scoring the mean bout length. The red stars signify the low p value of the Wilcoxon Rank Sum test. **(D)** IRR analysis confirmed that the grooming indices of the new user were comparable to that of the previous user (p = 0.4042, a: 0.1267, b: 1.0888, a_CI95: [-0.1416 0.3840], b_CI95: [1.0441 1.1355]).

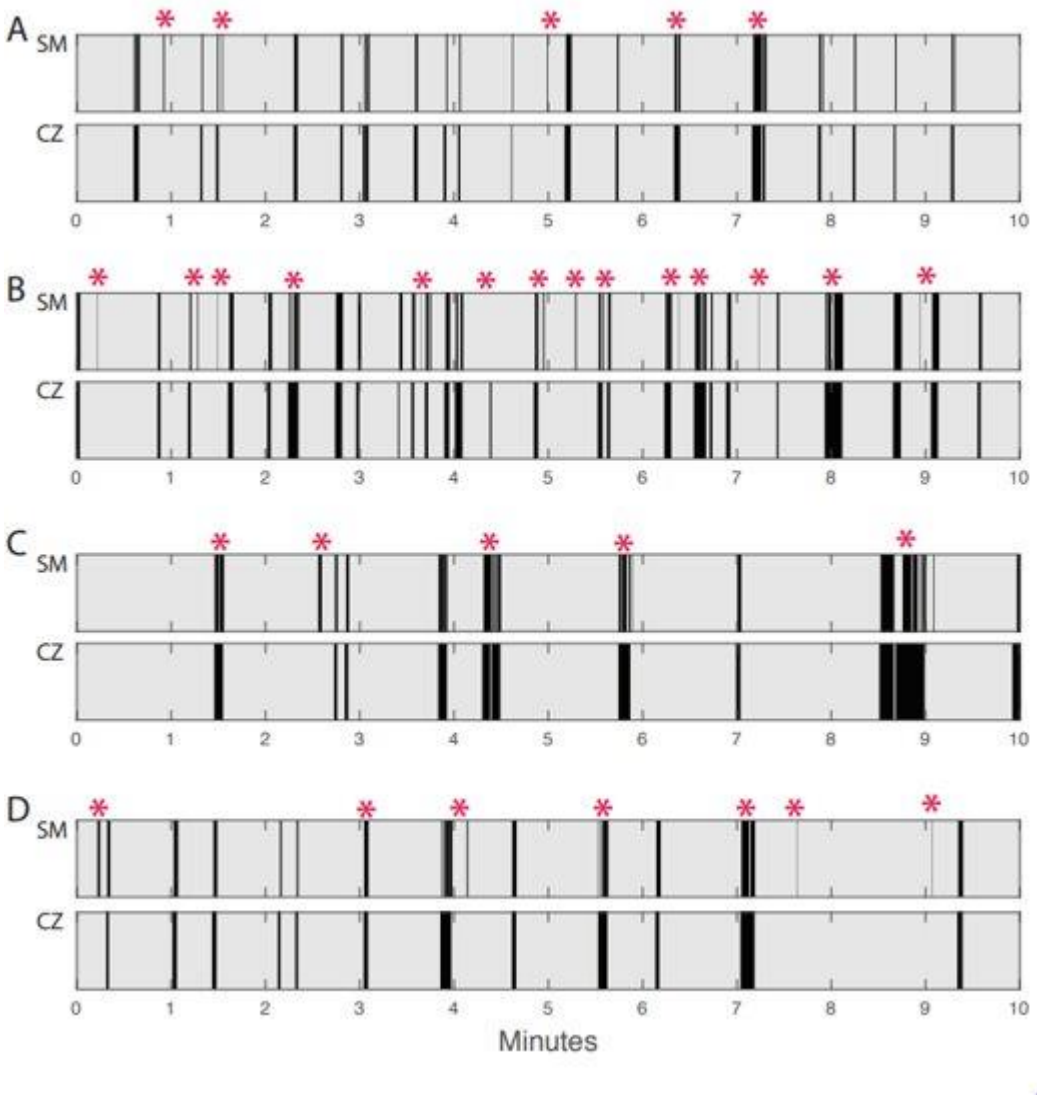


Figure 5: Series of Ethogram comparisons between the current and previous users depicting the variations in scoring of grooming bouts. A, B, C, and **D** represent the different videos where we observed marked differences in grooming between the two users. Each red asterisk signifies an instance where a grooming bout scored intact by the previous user was broken up by the current user.

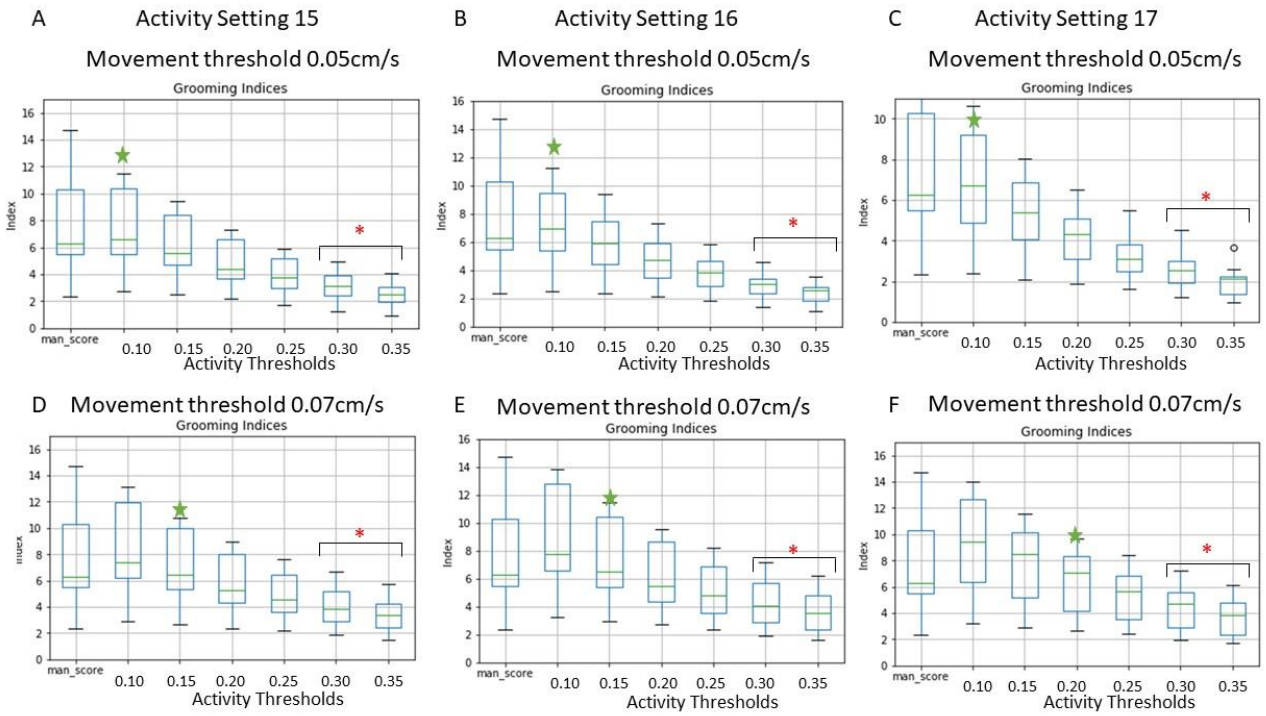


Figure 6: We used initial acquisition activity settings of 15, 16, and 17 to generate grooming indices at movement thresholds of 0.05 cm/s and 0.07 cm/s over an activity threshold range of 10 to 35% (0.10 to 0.35 in increments of 0.5). The green stars represent the parameter set generating the least statistically significant differences between manual and automated scores and the red asterisks represent the parameter sets generating differences with high statistical significance. **(A), (B), (C)** are grooming indices generated at a movement threshold of 0.05 cm/s over an activity threshold range from 0.10 to 0.35 (with 0.05 increments) for each of the activity settings. **(D), (E), and (F)** are grooming indices generated at a movement threshold of 0.07 cm/s over an activity threshold range from 0.10 to 0.35 (with 0.05 increments) for each of the activity settings.

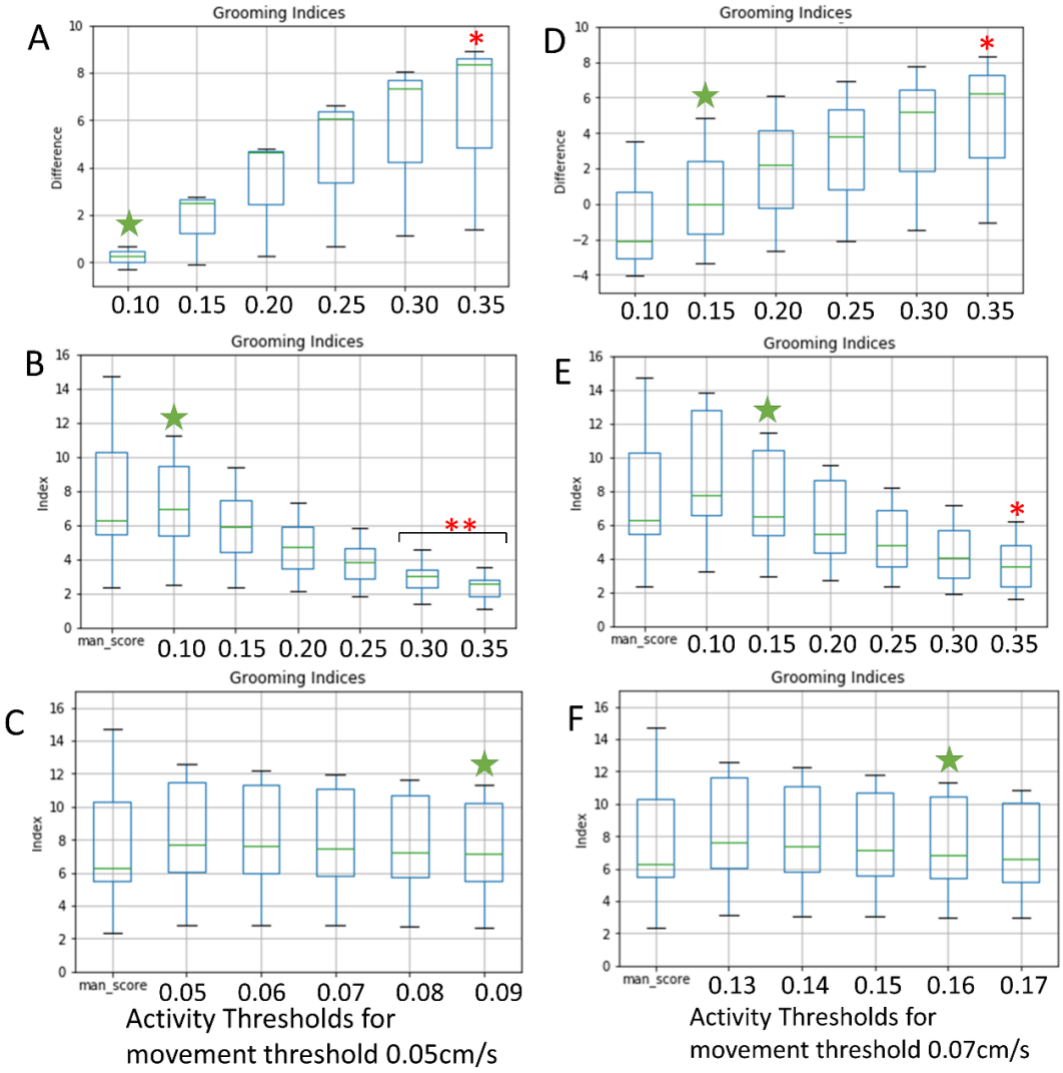


Figure 7: At an initial acquisiton activity setting of 16, the working subset of ten videos when implemented with the different combinations of parameter settings showed the least statistically significant differences between manual and automated scores at a movement threshold of 0.05 cm/s and an activity threshold of 10%(0.10). The green stars represent the parameter sets with least statistically significant scores and the red asterisks represent parameter sets with the most significant p values. **(A)** Box plot of differences between manual and automated scores generated at movement threshold 0.05 cm/s and activity threshold range between 0.10 to 0.35 in 0.05 increments. **(B)** Similar box plot comparisons between automated and manual grooming indices. **(C)** When closely focused at an activity threshold range from 0.05 to 0.09, we saw that an activity threshold of 0.09 yields the least statistically significant results **(D)** At a movement threshold of 0.07 cm/s, the activity threshold of 0.15 yields automated scores that are not statistically significant from manual scores **(E)** The results from **C** were supported by the box plots showing the manual scores in comparison to automated scores where we see how the combination of 0.05 and 0.15 yields the least differences. **(F)** When focused at an entire range from 0.13 to 0.17, an activity threshold of 0.16 gives automated scores that are the closest to manual scores.

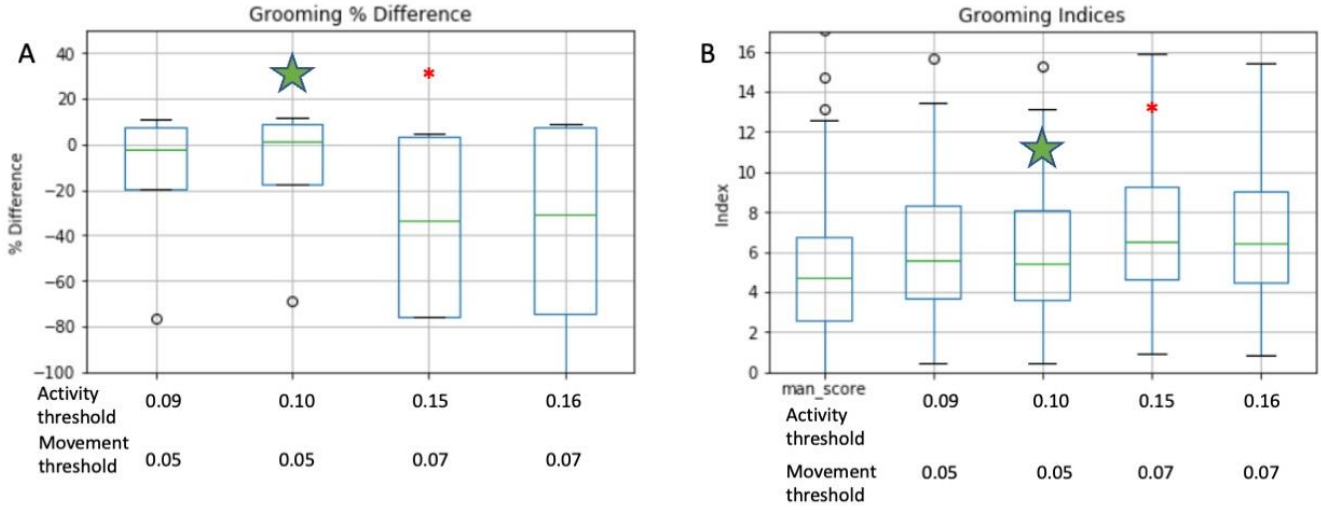


Figure 8: The combination of a movement threshold of 0.05 cm/s and an activity threshold of 10% (0.10) when applied to the entire set of 50 animals yields manual scores that do not show statistically significant differences between automated and manual scores. (A) Out of the four settings, a combination of 0.05 cm/s and 10% (0.10) yields the highest Mann Whitney p value 0.276 and shows differences between manual and automated scores approaching zero **(B)** When plotted against the manual grooming indices, the same combination of 0.05 cm/s movement threshold and 10% (0.10) activity threshold yielded scores that had a median closet to the manual scores.

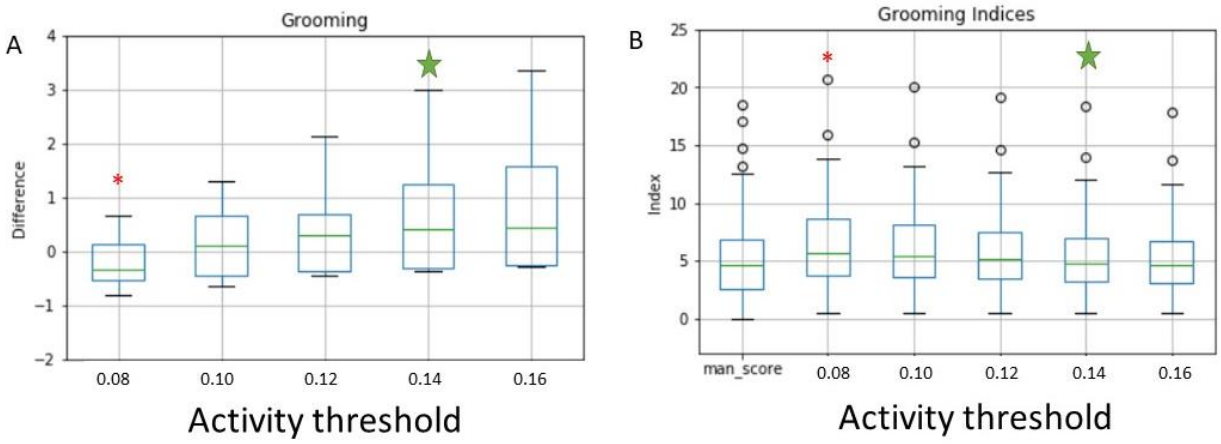


Figure 9: When an entire range of activity thresholds from 0.08 to 0.16 was focused on for the entire set of 50 animals, the combination of a movement threshold of 0.05 cm/s and an activity threshold of 10% (0.10) shows the least statistically significant differences between automated and manual scores. However, a combination of movement threshold of 0.05 cm/s and activity threshold of 14% (0.14) yielded the best set of grooming indices with medians closer to manual scores. (A) Out of the four settings, a combination of 0.05 cm/s and 10% (0.10) shows differences between manual and automated scores approaching zero **(B)** When plotted against the manual grooming indices, the combination of 0.05 cm/s movement threshold and 14% (0.14) activity threshold yielded scores that had a median closest to the manual scores.

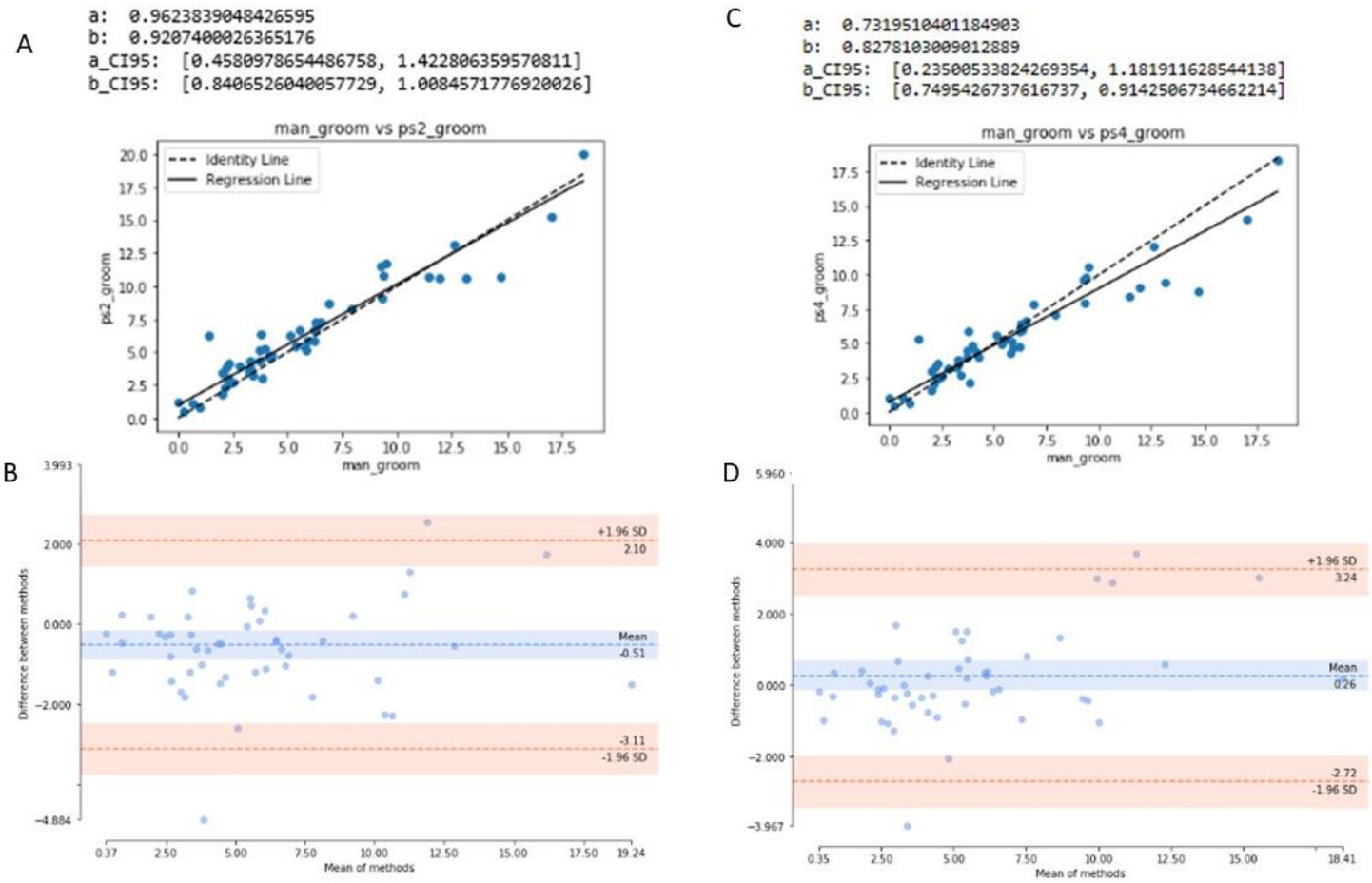
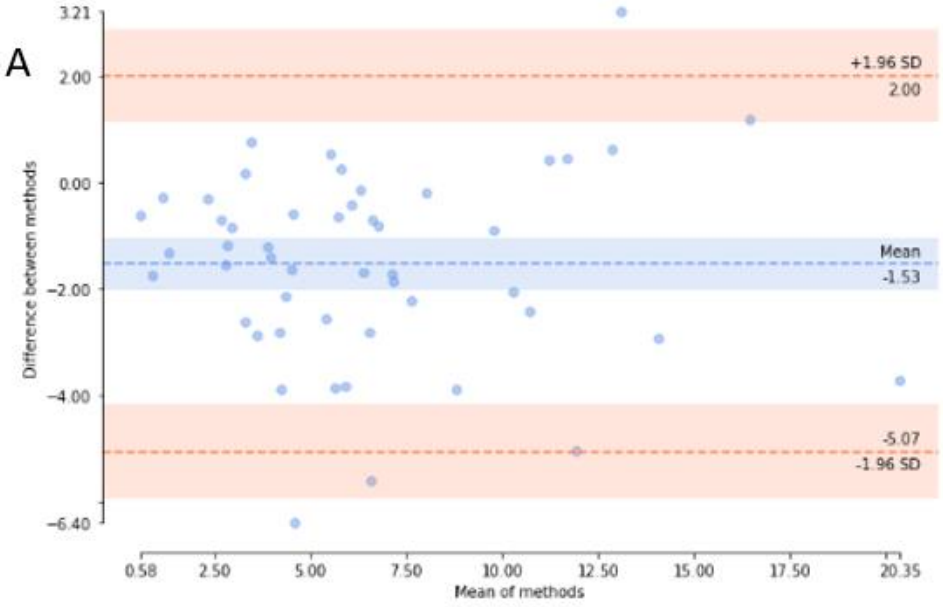


Figure 10: OLP regression and Bland Altman plots of the manual scores against the automated scores generated at the two identified software settings that vary at the individual animal and the population level. (A) OLP regression plot of the automated scores (grooming indices) generated at a movement threshold of 0.05 cm/s and activity threshold of 10% (0.10) against the manual scores. **(B)** Bland Altman plot of the means of the automated methods against the differences between the methods for the movement threshold of 0.05 cm/s and activity threshold of 10% (0.10) **(C)** OLP regression plot of the automated scores (grooming indices) generated at a movement threshold of 0.05 cm/s and activity threshold of 14% (0.14) against the manual scores. **(D)** OLP regression plot of the automated scores (grooming indices) generated at a movement threshold of 0.05 cm/s and activity threshold of 14% (0.14) against the manual scores.



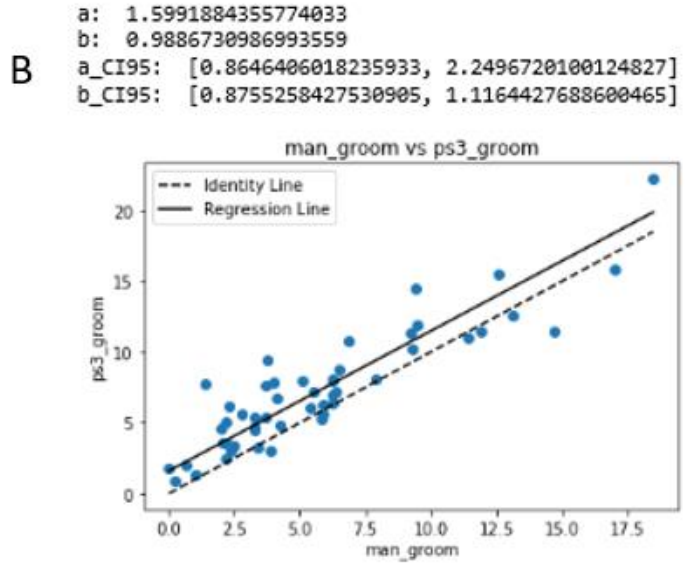


Figure 11: OLP and Bland Altman Plot of the manual scores against the automated scores generated at a movement threshold of 0.07 cm/s an activity threshold of 0.15% for the individual population. (A) OLP regression plot of the automated scores (grooming indices) generated at a movement threshold of 0.07 cm/s and activity threshold of 15% (0.15) against the manual scores.**(B)** Bland Altman plot of the means of the automated methods against the differences between manual and EthoVision scores for the movement threshold of 0.05 cm/s and activity threshold of 15% (0.15).

References

Aimon S. and I. C. Grunwald Kadow, 2019 Studying complex brain dynamics using *Drosophila*. J. Neurogenet. 171-177.

Andrew D. A., M. E. Moe, D. Chen, J. A. Tello, R. L. Doser, *et al.*, 2020 Spontaneous motor-behavior abnormalities in two *Drosophila* models of neurodevelopmental disorders. J. Neurogenet. https://doi.org/10.1080/01677063.2020.1833005

Bear M. F., Huber K. M., and Warren S. T., 2004 The mGluR theory of fragile X mental retardation. Trends Neurosci. 27:370–377.

Boos D., and L. A. Stefanski, 2011 P-Value Precision and Reproducibility Am. Stat. 65:213-221. doi: 10.1198/tas.2011.10129. Epub 2012 Jan 24.

EthoVision XT. Noldus Information Technologies, Wageningen, The Netherlands. EthoVision XT - Video tracking software | Noldus

Coll-Tane M., A. Krebbers, A. Castells-Nobau, C. Zweier, A. Schenck, *et al.*, 2019 Intellectual disability and autism spectrum disorders 'on the fly': insights from *Drosophila*. Dis. Model. Mech. 12:dmm039180.

Dockendorff T. C., H. S. Su, S. McBride, Z. Yang, C. H. Choi *et al.*, 2002 Drosophila lacking dfmr1 activity show defects in circadian output and fail to maintain courtship interest. Neuron. 34:973-84.

Feldbauer M.J., 2020 Automated analysis of grooming behavior in Drosophila using EthoVision XT. Unpublished. Lycoming College Departmental Honors Thesis.

Garber K.B., J. Visootsak, and S.T. Warren, 2008 Fragile X Syndrome. Eur. J. Hum. Genet. 16(6):666-672.

Giavarina D., 2015 Understanding Bland-Altman analysis. Biochem. Med. (Zagreb). 25:141-152.

Guo S., S. Zhong, and A. Zhang, 2013 Privacy-preserving Kruskal-Wallis test. Comput. Methods. Programs Biomed. 112:135-145.

Hagedorn, J., and J. Hailpern, 2008 VCode and VData: Illustrating a new framework for supporting the video annotation workflow. Paper presented at the Working Conference on Advanced Visual Interfaces: AVI 08, Naples, Italy.

Hagedorn, J., and J. Hailpern, 2008. VCode. http://social.cs.uiuc.edu/projects/vcode.html

Hampel S., C. E. Mckeller, J. H. Simpson, and A. M. Seeds, 2017 Simultaneous activation of parallel sensory pathways promotes a grooming sequence in Drosophila. eLife, 6, e28804.

Hampel S., K. Eichler, D. Yamada, D. D. Bock, A. Kamikouchie, and A. M., Seeds, 2020 Distinct subpopulations of mechanosensory chordotonal organ neurons elicit grooming of the fruit fly antennae. eLife9:e59976.

Hamada F.N., M. Rosenzweig, K. Kang, S.R. Pulver, A. Ghezzi, *et al.*, 2008 An internal thermal sensor controlling temperature preference in Drosophila. Nature 454:217-220.

Hannum C.L., 2017 Genetic analysis of spontaneous grooming behavior in the fruit fly *Drosophila melanogaster*. Unpublished. Lycoming College Departmental Honors Thesis.

Hazra A., and N. Gogtay, 2016 Biostatistics series module 3: comparing groups: numerical variables. Indian J. Dermatol. 61:251-260.

Huang W., A. Massouras, Y. Inoue, J. Peiffer, M. Ramia *et al.*, 2014 Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome. Res. 24:1193-208.

Inlow J. K., and L. Restifo, 2004 Molecular and comparative genetics of mental retardation. Genetics. 166:835-881.

Irwin S. A, B. Patel, M. Idupulapati, *et al.*, 2001 Abnormal dendritic spine characteristics in the temporal and visual cortices of patients with fragile-X syndrome: a quantitative examination. *Am J Med Genet.* 98:161–167.

Kain, J., C. Stokes, Q. Gaudry, X. Song, J. Foley, *et al*., 2013 Leg-tracking and automated behavioural classification in Drosophila. Nat. Commun. 4(1910).

Kaufmann W. E., H. W. Moser., 2000 Dendritic anomalies in disorders associated with mental retardation. *Cereb Cortex.* 10:981–991.

Kaur K., A. F. Simone, V. Chauhan, and A. Chauhan, 2015 Effect of bisphenol A on Drosophila melanogaster behavior—a new model for the studies on neurodevelopmental disorders. Behav. Brain. Res. 284:77-84.

Ludbrook J., 1997 Comparing methods of measurement. Clin. Exp. Pharmacol. 24:193-203.

Ludbrook J., 2010 Linear regression analysis for comparing two measurers or methods of measurement: but which regression? Clin. Exp. Pharmacol. Physiol. 37:692-9.

Mackay T. F. C., E. A. Stone, and J. F. Ayroles, 2009. The genetics of quantitative traits: challenges and prospects. Nat. Rev. Genetic. 10:565-77.

Mackay T. F.C., S. Richards, E.A. Stone, A. Barbadilla, J. F. Ayroles, *et al.*, 2012 The *Drosophila melanogaster* genetic reference panel. Nature. 482:173-178.

McDiarmid T. A., A. J. Yu, and C. H. Rankin, 2017 Beyond the response—High throughput behavioral analyses to link genome to phenome in *Caenorhabditis elegans*. Genes Brain Behav. e12437.

Michel C. I., R. Kraft, and L. L. Restifo, 2004 Defective neuronal development in the mushroom bodies of Drosophila fragile X mental retardation 1 mutants. J. Neurosci. 24:5798-809.

Nakamoto M., V. Nalavadi, M. P. Epstein, U. Narayanan, G. J. Bassell, S. T. Warren, 2007 Fragile X mental retardation protein deficiency leads to excessive mGluR5-dependent internalization of AMPA receptors. Proc. Natl. Acad. Sci. USA. 104:15537–15542.

Pandey U. B., and C. D. Nichols, 2011 Human disease models in *Drosophila melanogaster* and the role of the fly in therapeutic drug discovery. Pharmacol. Rev. 63:411-436.

Perry D.C., X. L. Griffin, M. Dritsaki, M. L. Costa, and N. Parsons, 2017 Becoming confident about confidence intervals. Bone Joint J. 99B:563-565.

Pfeiffer B. D., A. Jennet, A. S. Hammonds, T. B. Ngo, S. Misra, *et al.*, 2008 Tools for neuroanatomy and neurogenetics in Drosophila. Proc. Natl. Acad. Sci. USA. 105: 9715-9720.

Phillis R. W., A. T. Bramlage, C. Wotus, A. Whittaker, L. S. Gramates, *et. al.*, 1993 Isolation of mutations affecting neural circuitry required for grooming behavior in Drosophila melanogaster. Genetics. 133:581-592.

Qiao B., C. Li, V.W. Allen, M. Shirasu-Hiza, and S. Syed, 2018 Automated analysis of long-term grooming behavior in *Drosophila* using a *k*-nearest neighbors classifier. elife. 20147:e34497.

Restifo L. L., 2005 Mental retardation genes in *Drosophlia*: new approaches to understanding and treating developmental brain disorders. Ment. Retard. Dev. Disabil. Res. Rev. 11:286-294.

Rosner B. and D. Grove, 1999 Use of Mann-Whitney U-test for clustered data *Stat. Med.* 18:1387-1400.

Sachs B. D., 1988 The development of grooming and its expression in adult animals. Annals of the New York Academy of Sciences. 525:1–17.

Seeds A.M., P. Ravbar, P. Chung, S. Hampel, F. M. Midgley Jr. *et al.*, 2014 A suppression hierarchy among competing motor programs drives sequential grooming in Drosophila. elife. 3: e02951.

Sokolowski M. B., 2001 *Drosphila*: Genetics meets behaviour. Nat. Rev. Genet. 2:879-890.

Spechhia V., S. D'Attis, A. Puricells, and M. P. Bozzetti, 2017 dFmr1 plays roles in small RNA pathways of *Drosophila melanogaster*. Int. J. Mol. Sci. 18:1066.

Sprujit B. M., J. A. R. A. M. van Hoof, and W. H. Gispen, 1992 Ethology and neurobiology of grooming behavior. Physiol. Rev. 72:825–52.

Szebenyi L., and S. J. Andrew, 1969 Cleaning behaviour in *Drosophila melanogaster*. Anim. Behav. 4:641-651.

Yanagawa, A., W. Huang, A. Yamamoto, A. Wada-Katsumata, C. Schal, *et al.*, 2020 Genetic basis of natural variation in spontaneous grooming in *Drosophila melanogaster*. G3. 10:3453-3460.