Automated Analysis of Grooming Behavior in *Drosophila* Using EthoVision XT

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

By
Mikayla J. Feldbauer
Lycoming College
May 2020

Approved by:

Dr. Andrew Brandon

CHRISS MCDONALD (May 5, 2020)

Dr. Chriss McDonald

Dr. Mary Morrison

Dr. David Andrew

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

Table of Contents

Abstract

Studying behavior in the fruit fly *Drosophila melanogaster* can prove beneficial for better understanding several human diseases and potentially developing treatments to combat them. For example, the repetitive behaviors associated with mental disorders in those with Fragile X Syndrome can manifest in the equivalent fly models as excessive grooming behavior. Though there have been some published automated analyses, the typical route in order to study these behaviors in flies is for a human to watch hours of video and manually score each fly for the behaviors of interest. This is a time-consuming and inefficient process. In this study, we present a novel automated analysis of grooming behavior in *Drosophila* using the behavior tracking software EthoVision XT. Preliminary results suggest that the extraction of distinct behaviors, including grooming, from videos of spontaneously behaving *Drosophila* is indeed possible. Future studies should focus on further refinement of the EthoVision parameters needed to accurately distinguish between exclusive behaviors.

Background

The fruit fly *Drosophila melanogaster* is a well-established model organism for studying human disorders for many reasons, including the ability to raise genetically identical populations in a controlled environment and the large number of relevant conserved genes between humans and flies (Mackay and Anholt 2006). By studying behaviors in mutant and wild-type flies scientists can understand genetic aspects of human disorders associated with similar behaviors. For instance, humans with the neurodevelopmental condition Fragile X Syndrome (FXS) often show stereotyped repetitive behaviors similar to those seen in Autism Spectrum Disorder (ASD) (Garber *et al.* 2008; American Psychiatric Association 2013), and genetic fly models of FXS replicate these repetitive behaviors as an extremely elevated level of patterned grooming movements (McBride *et al.* 2013; Tauber *et al.* 2011; Andrew *et al.* in prep). Many behaviors in both humans and flies, however, are not caused by distinct genetic variants in one or two genes, but rather are quantitative traits that result from the influence of many genetic variants *and* environmental factors. Uncovering the genetic associations of quantitative traits in humans is an arduous task, but these associations are more genetically tractable in flies. This is due in part to the experimental control available in this model system, including the ability to control for genetic backgrounds and environmental conditions, the existence of a well-annotated genome, and the ability to create and readily evaluate mutations (Mackay *et al.* 2009; Hales *et al.* 2015).

In order to understand the genetic regulation of behaviors it is crucial to determine which genetic locations, or loci, are associated with quantitative traits relating to that behavior. One way of discovering which genetic loci influence quantitative traits is through genome wide association studies (GWAS), which correlate genomic data with quantifiable behaviors to draw associations of distinct loci with those behaviors. In flies, researchers utilize a series of

genetically distinct populations of animals whose environmental conditions can be easily and precisely regulated and examine variation in behavioral metrics between these populations (Mackay *et al.* 2009). The *Drosophila* Genetic Reference Panel (DGRP) was created to facilitate these types of GWAS analyses (Mackay *et al.* 2012). The DGRP is a community resource that contains just under 200 genetically distinct and isogenized fly lines whose entire genomes have been completely sequenced and made available (Mackay *et al.* 2012; Huang *et al.* 2014). To perform GWAS studies using the DGRP, researchers must first quantify their phenotype(s) or behavior(s) of interest in as many DGRP lines as possible and submit those values to the DGRP webtool for analysis (http://dgrp2.gnets.ncsu.edu; Mackay *et al.* 2012; Huang *et al.* 2014). Previously, the Andrew lab has analyzed around 40 *Drosophila* lines from the DGRP for variation in grooming behavior using a labor-intensive manual scoring procedure (Hannum 2017). The goal of the current project is to streamline and automate the process of behavior annotation by utilizing a video analysis software package called EthoVision XT® (hereafter referred to as EthoVision; Noldus Information Technologies, Wageningen, The Netherlands). This software was originally designed to automate the detection of several different behaviors in rodents, but should be sufficiently adaptable to decipher grooming motions from recorded videos of behaving flies, which is how we propose to utilize it.

Previously, to determine the number and duration of grooming bouts, videos of *Drosophila* flies would be recorded and manually scored for different behaviors during each frame of a video. The manual scoring of videos is a tedious and time-consuming process; for instance it could take longer than 20 minutes to score a 10-minute long video for a fly that exhibits many changes between behaviors. Automation of this process with EthoVision aims to save researchers time and resources while also enabling a higher throughput and standardization

of behavioral analysis. A previous study performed in the Andrew lab (McLaughlin 2018) concluded that EthoVision can be used to accurately determine the walking behavior of flies in a manner that is comparable to manually-scored results. The objective of this current project is to determine if EthoVision can likewise be used to accurately quantify grooming behavior in flies. As grooming is a much more nuanced and variable behavior than walking, with wide variation in behaviors both within and between animals, this is a much more difficult task to do with this or any piece of software.

Drosophila flies, like all limbed animals, regularly perform grooming motions in order to rid their bodies and sensory receptors of dust particles (Szebenyi 1969; Dawkins & Dawkins 1976). In flies, grooming is characterized as a group of stereotyped movements where the fly rubs one or more of its legs together or on another part of its body to remove debris (Szebenyi 1969). Broad sweeping of the legs over the body can cause a dust particle to make its way to the bristles of the legs, where a subsequent rubbing of the legs together causes the particle to completely drop off the legs (Szebenyi 1969). While flies groom naturally and spontaneously, an increase in grooming can be seen after applying dust to the animal, after which the fly will groom until clean in a stereotyped pattern of anterior to posterior progression (Phillis *et al.* 1993; Seeds *et al.* 2014).

Some of the very first studies performed on grooming required visual tracking of the behavior in real-time. Szebenyi (1969) describes a process whereby the observer watches a fly groom (the actual fly, not a recording) and calls out symbols to distinguish the various types of grooming. These symbols are audio-recorded so that a written record can be made later on for subsequent analysis (Szebenyi 1969). A later study by Dawkins and Dawkins (1976) also included the real-time observation of a blowfly, though the grooming bouts were recorded using

a keyboard/computer system that printed the record on paper. These first few studies on grooming laid the groundwork for how and why to systematically study grooming as a behavior. However, as they often required real-time manual observation, they were not very efficient and relied heavily on extensive training and expertise. Automated analysis, which is proposed in this study, allows for higher throughput, more accurate records, and increased repeatability between experiments.

Previous studies have looked at automating the analysis of various behaviors in *Drosophila* (Ramazani *et al.* 2007), with some focusing explicitly on grooming. More recent approaches employ machine learning to distinguish one behavior from another (Kain *et al.* 2013; Qiao *et al.* 2018). Ramazani *et al.* (2007) created an analysis pipeline for the automated detection of several exploratory behaviors in *Drosophila* as well as in mice. They used custom Perl scripts to analyze images by subtracting the background so that only the flies that moved were left (Ramazani *et al.* 2007). This approach was shown to be effective in analyzing exploratory behavior for one or more flies in an arena, but is extremely limited in its ability to differentiate stereotyped, limb-focused behaviors like grooming.

Kain *et al.* (2013) tracked individual legs on their flies by tethering the fly, placing dye on the legs, and using a laser to track the legs from below. The tracking of individual legs allows for the study of intricate behavior such as grooming and gait analysis. Kain *et al.* (2013) used a type of machine learning algorithm called *k*-nearest neighbors, which uses a training data set to categorize unknown data points, to classify the frames of video that they recorded. To benchmark their data, two independent human scorers labeled videos for a total of 5 flies for 12 different behaviors. These results were used as the training set for their algorithm and the agreement between both human scorers (71%) was used as the benchmark for algorithm

performance (Kain *et al.* 2013). Using this machine learning method, Kain *et al.* (2013) achieved an accuracy of 66%, only 5% lower than the accuracy of the human scorers, though the interrater reliability between their human scorers (71%) was already a low value. This approach is technically demanding both in regard to the significant manipulations of the animal for recording (e.g. they artificially tether their fly to a fixed position, which precludes spontaneous motion) as well as the hardware and software requirements, which would be available to only well-funded and highly-trained laboratories.

Qiao *et al.* (2018) more recently developed a method that allows for long-term tracking of *Drosophila* and an automated analysis of grooming. They developed this method in response to the limitations posed by other automated analyses that were only suited to shorter tracking times and/or fewer animals. Similar to Kain *et al.* (2013), Qiao *et al.* (2018) used a *k*-nearest neighbor machine learning algorithm to categorize the video frames as grooming, locomotion, or rest. Their algorithm classified grooming with over 90% accuracy when compared to their manual scores (Qiao *et al.* 2018), but still relies heavily on precise recording conditions and custom, difficult to implement Matlab scripts.

Although there are already published methods for detecting behaviors in flies, including grooming, all prior published methods require extensive computational power, very specific monitoring systems, elaborate software implementation, and are consequently resource intensive. This current study presents a novel automated method for behavior detection in flies with a focus on grooming that utilized a commercially available software package, EthoVision, to extract behaviors from previously recorded videos of *Drosophila.*

Our approach differs from these previous studies in that it is not as reliant on intensive recording setups or technically challenging computational expertise and is therefore more

accessible to those who might not be familiar with the level of programming needed to implement a machine learning analysis. EthoVision does the video analysis and even outputs statistics that the user can utilize such as time spent in a certain behavior and the percentage of time spent in that behavior. After the parameters are set and validated, the only computation required to perform future analyses would all occur through the EthoVision interface, with which the user interacts graphically rather than through code. This also saves on computational resources as the only commercially required software is EthoVision, no special set-up is needed for recording the flies, and a massive amount of computing power is unnecessary for implementation. Additionally, our lab has a large collection (over 150 videos) of previously scored videos against which any automated analyses can be compared.

Materials and Methods

Establishing a collection of Drosophila *behavior videos*

Previous members of the lab established a collection of 157 videos that captured spontaneous *Drosophila* behavior in more than 750 flies from 34 lines of the DGRP (Hannum 2017). For each video, up to six flies at a time were manually aspirated into individual wells of a 96-well plate and covered with a clear glass slide to create a behavioral arena where flies would spontaneously behave in isolation for 10-minute intervals. The wells featured in the videos are arranged in two rows of three; the distance across one row of three wells is 2.6 cm (**Figure 1**). Each video was cropped to 600 seconds (10 min) in length after recording using Adobe After Effects (Adobe After Effects CC 2014; Adobe Inc. San Jose, CA). The DGRP line, sex, and well number was recorded for each fly and is recorded in a spreadsheet along with the fly's manually-scored grooming index (Hannum 2017).

Figure 1. Three wells of the 6 per video are shown – note the third well is vacant. The distance across one row is 2.6 cm. Videos with 6 wells would have another row below the one shown.

Establishing a reference set of manually-scored videos

Select videos were chosen from the lab's collection of previously-scored videos and rescored at the initiation of this current study for grooming, standing, walking, and falling behaviors in order to familiarize myself with the vagaries of fly behaviors and establish a reference set with which to compare the EthoVision results obtained in this study. We chose lines with a wide variation in grooming index (GI), which is defined as the total time the fly spends grooming divided by the total number of seconds that the animal was observed (i.e. 600 seconds). These videos had previously been manually scored (Hannum 2017) using the VCode video annotation software (Hagedorn *et al.* 2008). During manual scoring with VCode, the observer watches the video and notes when the fly performs a specific behavior and records that behavior with coded keystrokes. For instance, if the scorer is keeping track of grooming, every time the fly begins a grooming bout, they would press the button on their keyboard that corresponds to that fly's well in the video. They would then press that same button again once the grooming bout is finished. Several relevant metrics were calculated and recorded from the results of manual curation, such as the number and duration of grooming, falling, standing, and walking bouts. These manually scored indices from my current analysis were compared to the

indices calculated during a previous honors project for consistency and reliability between individual scorers across different projects (a total of 3 human scorers throughout 2 previous honors projects: Hannum 2017; McLaughlin 2018).

Determination of behaviors using EthoVision

The first step of automated scoring is to load each video into the EthoVision software and establish the proper software settings for each particular video. The arenas for each fly were defined for a video by drawing a circle over that arena. The video's scale was then determined by tracing the distance over three arenas within a single row and setting the distance to 2.6 cm as shown in **Figure 1**. After the arenas were defined for each fly in a video, the detection settings were tuned to isolate the dark fly from the light-colored background and define the limits and size of the animal for the program. To do this, we used EthoVision's automated setup option, wherein a box is drawn around each fly, avoiding the legs and wings as much as possible, after which the "advanced detection settings" were set to fine-tune the detection of the fly's body. These settings specify the gray scale values used to differentiate the fly's pixels (dark pixels with low grayscale values, minimum 0) from those of the arena's background (light pixels with high grayscale values, maximum 255). The activity settings for the flies are set such that the movement of the fly's legs is captured without much background interference being recorded. This is accomplished by fine tuning the activity threshold used by EthoVision to calculate the pixels that are changing in the background, which is measured as a difference in grayscale values from one frame to the next. I watched the pixels being recognized without the video in order to determine when the legs were captured while changes in the background were not, which insured that leg movements were accurately documented. When running experiments using EthoVision, the fist step is to do this setup for each new video being analyzed. This process will take care of

any lighting or other recording differences between videos, though the videos we record are largely uniform.

For the initial validation experiments, only certain flies were chosen from each video, corresponding to the manually-scored reference set mentioned above. This means that EthoVision only analyzed 1-4 flies for each video, while the other flies in the video were ignored. This was done because these initial trials were for benchmarking purposes and a more thorough analysis of the lab's manually scored video collection was to be done after the threshold values for the parameters were refined. This saved both computational time and manual time in the beginning stages of the project. Each of the flies chosen for this initial analysis also had their manual standing, falling, and walking indices recorded in addition to their previously recorded grooming index, pieces of information not present for the other flies in the lab's collection.

After the user specifies the detection settings for each video and puts all of the videos to be analyzed into a trial list, EthoVision then analyzes the videos and collects the data specified for each one and produces annotated video tracks for further analysis by the user. EthoVision's data profiles were utilized to bin each frame of the video into a behavior using the behavior-calling logic outlined in **Figure 2**. We binned behaviors into three categories of **walking**, **standing**, and **grooming**, each of which is mutually exclusive to the other, that is the fly can only exhibit one of these behaviors at a time. To determine which behavior a frame is classified as, the frame needed to meet three requirements for the initial analysis. These requirements consisted of falling into threshold values for EthoVision's *movement*, *mobility state*, and *activity state* parameters, initially characterized as shown in **Figure 2**. The logic behind our initial parameter settings is as follows: 1) standing flies are immobile and not moving their appendages,

2) walking flies are actively moving around the arena so their body outlines are mobile, and 3) the body of grooming flies is relatively constantly placed within an arena, but one or more appendages are rapidly moving to perform the grooming behavior. We reasoned that with this logic we could use the *movement*, *mobility state,* and *activity state* parameters of EthoVision (explained below) to differentiate these behaviors.

The first consideration of our initial logic is to determine the *movement* of each fly in an arena. The ***movement*** parameter is calculated based on the velocity of the center point of the animal, which EthoVision calculates by a center-weighted average of the pixels identified as the main body of the animal (see the red boxes on the fly in each panel of Figure 3). An animal is moving if its average velocity exceeds the start velocity parameter defined by the user, otherwise it is classified as not moving. For our initial trials, I chose the velocity threshold by both watching the integrated visualization that EthoVision provides and determining the velocity of the fly during different behaviors and looking to a previous honors project studying walking (McLaughlin 2018). For future studies, the threshold that is found to be the most accurate at calling all of the flies' behaviors should be used. We reasoned that animals that are walking should have a high moving classification, while those that are both standing and grooming should have a movement value below some threshold to be determined empirically.

The next component of our behavior classification logic considers the ***activity state*** parameter, which is calculated by comparing the pixels in the entire arena from one frame to the next and determining the percentage of those pixels that changed. There are only two available activity states: highly active and inactive. The pixels in yellow in **Figure 3**A are detected as the animal, while the pixels in purple are the pixels in the arena that changed from the previous frame. The pixels that changed during a grooming bout can be seen in **Figure 3**B. We reasoned

that we could differentiate grooming from standing using the *activity state* parameter in flies that are classified as "not moving" per the *movement* parameter (**Figure 2**).

Finally, to further differentiate grooming from standing and/or walking, we employed the ***mobility state*** variable (**Figure 2**), which is calculated by comparing all the pixels defined as the subject in one frame to the pixels defined as the subject in the subsequent frame. The percent difference between the pixel values is used to calculate *mobility state*, which ranges from 0 (no difference in pixels – stationary object) to 100 (every pixel moved – mobile subject). **Figure 4** illustrates the *mobility state* variable as the subject fly can be seen in one frame compared to the subsequent frame of the video (**Figure 4**A). An overlay of the two frames can be seen in **Figure 4**B. Note the difference in the position of the fly's body. EthoVision uses the overlap to calculate percent mobility, which is compared to the user-defined threshold values to calculate *mobility state*. There were initially three mobility states in our analysis: immobile, mobile, and highly mobile. The initial trials used the mobility variable. However, the results of those analyses showed that the use of the *mobility state* was unnecessarily restrictive – an observation that was confirmed after speaking with support representatives from Noldus. We therefore employed a simpler logic for subsequent trials, described later.

Sensitivity analysis framework for EthoVision threshold value settings

After the initial 13 manually-scored videos were loaded into EthoVision, the file was simply resaved under a different name for different parameter sets in order to avoid going through setup for each video every time a parameter set was changed. This allowed for the creation of different "experiments" where threshold values for two parameters were held constant, and the value of the remaining parameter was changed incrementally. An example would be holding the *movement* and *mobility state* threshold values constant while changing the

activity state threshold from 0.05% to 0.65% in increments of 0.1%. The data profiles were used to bin each frame of the video into a different behavior class (i.e. either walking, standing, or grooming), while the analysis profile enabled visualization of each called behavior against the actual video of that fly spontaneously behaving.

Data analysis of EthoVision statistics

For each parameter set, EthoVision outputs a statistics file. These statistics were composed of user-defined metrics, including the cumulative duration in seconds and the percentage of the total video time the fly spent in each behavior class. These files were exported as excel files for further analysis. The files were then put through a data analysis pipeline contained within custom made Python scripts using the data analysis package Pandas (McKinney 2010). The files were first converted to comma separated values (CSV) files for easier extraction using Python. The values contained within each file were stored in a Pandas DataFrame, which is a data structure similar to an excel table. The grooming, standing, and walking indices for each fly were calculated by looping through each behavior's DataFrame containing time spent in that behavior. Three-dimensional scatter plots were then created by plotting the grooming, standing, and walking indices on each of the three axes. The manually scored point for each fly was also plotted on the same plot for comparison of the manually scored indices to the EthoVision calculated indices. Each scatter plot contained points that represented each parameter set for a certain fly.

Box plots were also made as another comparison between the manually scored and EthoVision scored results. This was done by looping through the DataFrames to find the difference between the manually-scored and the EthoVision-called values for each fly; i.e. manually scored minus EthoVision called. The differences for each fly were used to create a

section of the box plot representing an entire parameter set. We chose to use the differences at first in order to gain an overview of how well each parameter set was performing. In subsequent analyses, once we refined our logic and threshold values, we plotted the distribution of raw values for each parameter set and compared those to the EthoVision results. These box plots were created for each of the behaviors in each of the main "experiments." Three main experiments were initially performed. In each of the experiments, the threshold values were held constant for two of the parameters and varied for the third. The threshold values that were changed for each experiment are outlined in **Table 1** (shown in red).

Creation of Python pipeline to streamline data analysis

After several rounds of data analysis, we sought to streamline the data analysis process by compiling the methods needed to organize and analyze the data into a class in Python. A "class" in object orientated programming languages (such as Python) allows the user to define a new object type. The class contains methods which are flexible (rather than being hardcoded) and, in this case, allow for more variation in the data entered. The methods contained within this data analysis class involve ones for the conversion of the excel files to CSV files, input of the data into Pandas data frames, calculation of the difference between manually scored and EthoVision calculated indices, and statistical analysis including creation of box plots and regression analysis.

Re-evaluation of initial classification scheme

Initial results showed that the logic used to group frames into behavior results was too restrictive and as a consequence EthoVision did not bin every frame of the video into one of our three behaviors. This was shown by comparing the total amount of time called for all mutually-exclusive behaviors to the total 600 recorded seconds for each fly – the logic *should* result in

every frame being called in one of the three potential behaviors. With our initial logic (**Figure 2**), however, each arena contained fewer than 600 seconds of behavior calls, some even containing fewer than 500 seconds (**Table 2**A). We observed that many frames (and therefore many seconds of video) were not being binned into any of the available behaviors. This was most likely due to the restrictions on each frame from the binning logic. The frames that were not being binned into a behavior result likely did not meet *all* of the requirements for that result. For example, if the fly is walking and meets the "moving" criteria but not the activity state requirement, it would not be binned into any result. After analyzing our preliminary results and discussing our implementation problems with the software manufacturer, we decided to simplify our behavior-calling logic to only use two of EthoVision's parameters: *movement* and *activity state*. This simplified logic reasons that the *movement* parameter will differentiate walking (moving) from both standing and grooming (both *not* moving), and a subsequent analysis of *activity state* will further divide the "*not* moving" behaviors into either standing (inactive) or grooming (highly active). Using this new logic (**Figure 5**), the total amount of time per arena tracked was much closer to the total 600 seconds per arena (**Table 2**B). The few seconds that are not tracked for each arena could be due to the averaging variable that EthoVision includes for each of its parameters. This is a measure of how many frames EthoVision averages over; for each of my parameters, this metric is set to 3 frames.

After the analysis of initial results, we therefore removed the mobility parameter from our EthoVision behavior-calling logic (**Figure 5**). This simpler logic is less restrictive and allows for more frames of every video to be binned into an appropriate behavior result.

The parameter sets for two experiments using this logic are shown in Table 6. For each of these experiments, the activity threshold was varied from 0.05% to 0.45% in increments of

0.05% while the movement threshold was held constant. The different independent variable between the two experiments is simply the movement threshold. There were three different movement thresholds, chosen based on the preliminary results that 0.07 cm/s was the most accurate threshold. A previous honors thesis (McLaughlin 2018) determined that this was indeed the most suitable threshold for walking and my results from last semester seemed to replicate that result. Varying the movement threshold between these new experiments was meant to validate that the 0.07 cm/s threshold was the most appropriate for grooming as well.

Statistical analyses

We performed three main statistical analyses on the results from our simplified logic. The first was an ordinary least products (OLP) regression (Ludbrook 1997). An OLP regression is useful when there is potential error in both measurements. For this reason, it is often used to align one method against another; it is also useful in detecting and assessing systematic biases between methods. In our analysis, both the manually-scored and the EthoVision-scored indices contain some amount of error as neither of them are a fixed, known value so we used an OLP regression rather than an ordinary least squares regression which assumes one measurement is known (Ludbrook 1997). I made the plots shown in Figures 10 through 12 with a custom-built Python script that was adapted from previously validated Matlab code (Trujillo-Ortiz, 2020; gmregress, version 1.7.0.0, MathWorks File Exchange).

In the output for our OLP regression, a is the slope and b is the y-intercept of the regression line, with a_CI95 and b_CI95 being the 95% confidence intervals for a and b, respectively. A *proportional* bias is present if a_CI95 does not include 1. This means that the slope differs significantly from that of the identity line, seen as the tilting of the regression line about one point (Ludbrook 1997). A *systemic* bias is present if b_CI95 does not include 0. In this

case, the intercept of the regression line differs from zero, meaning the entire line will be shifted up or down but still parallel to the identity line (if there are no other biases present) (Ludbrook 1997).

The second statistical analysis we performed was the creation of a Bland Altman plot (Altman & Bland 1983). This plot features a scatter plot of the values within the population where the y axis is the difference between the two methods and the x axis is the mean of the measurements. Altman and Bland recommend that you should find 95% of your data points within 2 standard deviations of the average difference (Giavarina 2015). My plots were made using the pyCompare module within Python (Pearce 2019, Version 1.4.1).

The third and final statistical analysis used in this approach is the Mann Whitney U test (Mann & Whitney 1947). This nonparametric test is used for large samples that are not normally distributed. In the evaluation of this pairwise test, each data point from one sample is compared to that of the second sample. The test's null hypothesis states that the two samples originate from the same population (Nachar 2008). This means that if the null hypothesis is rejected ($p < 0.05$), one can conclude that the two samples originate from different populations.

Results

Initial logic for determining behaviors was too restrictive

Based on the threshold values and defined binning logic (**Table 1**, **Figure 2**), EthoVision bins each frame of the video into a certain behavioral result, either walking, grooming, or standing. The resulting called behaviors can be visualized using EthoVision's integrated track visualization feature (**Figure 6**). The lighter, not tan areas show which frames are the ones called as that behavior. In **Figure 6**A, for example, activity and mobility states and movement are shown for a fly that is walking around the arena. This animal in this arena is exhibiting a high

activity and mobility along with movement. These three combined indicate that the animal is walking as it is actively moving around the arena. Its center point is moving (high mobility) along with the pixels in the arena (high activity), which indicates that the fly is not standing still. In **Figure 6**B, activity and mobility states and movement are shown for a fly during a grooming bout. This animal is highly active (high activity state), moderately mobile (mobile mobility state), and not moving (low movement parameter). The important distinguishing factor for animals that are not moving (i.e. ones that are either grooming or standing) is their activity state value. The pixels in the arena that are not defined as the animal are changing a lot for the animal, shown in **Figure 6**B, hence the high activity state reading and classification as grooming as opposed to standing. This means that the legs are moving which indicates grooming behavior. These validating features of the animal at the time the behavior is being called indicate that the behavior is called accurately in this instance.

The indices calculated from the EthoVision called times for each behavior were close to the manually scored and validated indices for some flies, but not all of them (**Figure 7**). This can be seen by comparing the black dot, representing the manually scored indices, to each of the colored dots, representing the indices for varying parameter sets (ps1–ps8). All of the graphs shown in **Figure 7** are from Experiment 2 (which varied mobility state parameter settings), except for panel C which is from Experiment 1 (which varied activity state parameters) (**Table 1**). Each of the dots plotted on the graphs in **Figure 7** represents the point in three-dimensional space where the grooming index (GI), standing index (SI), and walking index (WI) meet when plotting each of the indices on the x, y, and z axes, respectively. As seen in **Figure 7**A-C, the manually scored indices were closest to the indices from parameter set 1 or 2 (red or orange dots, respectively), which corresponds to an upper mobility threshold of 5% and 6%, respectively, for

graphs A and B and an inactivity threshold of 0.05% and 0.15%, respectively for graph C. Each of these plots represent the results for one fly, designated by the trial number (corresponding to which video) and arena the fly was in. **Figure 7**A-C represents promising results where the manually called indices were close to the EthoVision called ones. However, **Figure 7**D-F show results where the manually called indices were quite far from the EthoVision called indices, as indicated by the black dot being very far from the colored dots along several axes.

Box plots were also made showing the difference between the manually called and EthoVision called indices for all flies in our training set (n = 13; **Figure 8**). The spread of the box shows the range between the first and third quartiles, while the green line in the middle shows the median difference between EthoVision and manual scores for each metric. The upper and lower lines of the whiskers are positioned at a distance 1.5*(Q3 – Q1) from the edges of the box, whereas the circles show outliers in the data sets. The data points are calculated by subtracting the EthoVision indices from the manually scored ones, so a tight distribution of values close to zero signifies that the two indices are very close to each other and the parameter settings are doing a good job of accurately calling behaviors, which is our aim. Any negative values represent a threshold for which EthoVision overcalls that behavior, while a positive difference would indicate that EthoVision is under-calling that behavior.

Simplified behavior-calling logic more accurately determines behaviors

After re-thinking the logic for EthoVision's binning algorithm, we performed experiments to determine the appropriate movement and activity threshold at a population level, having removed the mobility threshold due to the unnecessary restrictions it places on the determination of the behavior indices (**Figure 5**). In addition, we expanded the training set to include 35 animals in order to obtain better population level estimates of behaviors. **Figure 9**

shows box plots that compare the manual scores with EthoVision scores of all scored flies in our expanded training set (n = 35). **Figure 9**A shows Experiment 4, where the movement threshold is 0.05 cm/s and **Figure 9**B shows Experiment 5, where the movement threshold is 0.07 cm/s. For each experiment, the activity threshold is varied from 0.05% to 0.45% in increments of 0.05%. Each box plot shows the manual scores first, followed by the EthoVision scored indices. Each box on an individual plot shows the spread of indices for that parameter set. The "whiskers" coming off from the main box are a distance of 1.5*(Q3-Q1) away from the box itself. Any dots shown outside this range are outliers. Some extreme outliers were left off of the plot in order to better visualize the spread of data that is shown; however, these points are included in all statistical analyses performed on this data set. The walking indices (the graphs to the far right for each row) are all the same for each parameter set because within each experiment, the activity threshold is changing (**Figure 9**). According to the logic used for binning video frames into behavior results (**Figure 5**), a change in the activity threshold will only affect the indices for grooming and standing as these are the only two results that are affected by activity. In both experiments, however, walking is consistently over-called by EthoVision (**Figure 9**). For **Figure 9**A, parameter sets 3 and 4 appear to be a good fit for standing (based solely on the box plots) while parameter set 1 appears to do better for grooming. This disconnect is not ideal; the ideal parameter set would be appropriate for *all* behaviors. For **Figure 9**B, parameter set 3 appears to represent all three behaviors well, but as discussed later on, there are nonetheless biases present in these data.

We proceeded to look at the data from all parameter sets on a fly-by-fly basis. Based on **Figure 9**, the spread of data from parameter sets 1 and 2 from Experiment 4 (with movement threshold of 0.05 cm/s; **Figure 9**A) and parameter set 3 from Experiment 5 (with movement

threshold of 0.07 cm/s; **Figure 9**B) were closest to the spread for the manual scores across all three behavior metrics. We then sought to determine differences of scores in individual flies by plotting the manually-scored metric for grooming along the *x* axis and the EthoVision scores along the *y* axis. A perfect correlation of these scores (i.e. $x = y$) would show a distribution of scores along the identity line, and deviations from that line would indicate potential biases in EthoVision's scoring. **Figure 10** shows the results of this analysis of the grooming scores for manual compared to the first parameter set in Experiment 4, where the movement threshold for these results is 0.05 cm/s and the activity threshold is 0.05%. In addition, **Figure 10**A shows an OLP regression line and **Figure 10**B shows a Bland Altman plot for each of these data (activity threshold of 0.05%). The OLP regression plot suggests that there is a systemic ($b\_CI95 \neq 0$) bias present (**Figure 10**A), which is confounded by an apparent positive correlation of points on the Bland Altman plot, which suggests a slight proportional bias of EthoVision under-calling grooming for animals that have high grooming indices (**Figure 10**B).

Similar plots from Experiment 4 with poorly fitting parameters can be seen in **Figure 11**, where the activity threshold is 0.45%. In this case, the plot demonstrates that EthoVision is consistently under-calling grooming behaviors (i.e. not accurately identifying bona fide grooming bouts). Both a proportional ($a\_CI95 \neq 1$) and a systemic ($b\_CI95 \neq 0$) bias are present, as seen by the OLP regression plot (**Figure 11**A). The systemic bias is also apparent in the obvious positive trend of points on the Bland Altman plot for this data set (**Figure 11**B).

Figure 12 shows OLP regression plots and Bland Altman plots for Experiment 5 where the movement threshold was 0.07 cm/s and the activity threshold was 0.15%. The first row shows grooming (**Figure 12**A,B), the second standing (**Figure 12**C, D), and the third walking (**Figure 12**E, F). The p-values for each of these parameter sets when compared to the manually-

scored set using a Mann Whitney U Test were all greater than 0.05 which suggests that the EthoVision scores did **not** significantly differ from the manual scores. Typically, one would look for statistically significant results, but in this case we want our samples to not be statistically significant because we want them to not differ from the manual scores in any significant way.

This parameter set represents a good fit that somewhat accurately calls the flies' behaviors. However, the OLP regression shows that there are still biases present in the EthoVision results. Both a proportional ($a\_CI95 \neq 1$) and a systemic ($b\_CI95 \neq 0$) bias are present in grooming measures (**Figure 12**A), while a slight systemic bias ($b\_CI95 \neq 0$) is present in both the standing and walking plots (**Figure 12** C and E).

Analysis roadblocks and recommendation for future studies

The pipeline for data analysis is now complete and ready to be applied to the larger set of 750+ flies that have already been manually scored for grooming (Hannum 2017). This large-scale benchmarking, with a large and well documented data set would provide more statistical power to the pairwise population measures shown above. This was the original goal of our project, to include all 750+ animals in an automated scoring experiment to verify the ability of EthoVision to assess these behaviors in a large data set and thereby increase analytical throughput of the Andrew lab. However, two major setbacks, both of which were beyond our control, delayed and prematurely ended our analyses. First, the computer on which the EthoVision and data analysis occurred suffered from a hard drive wipe that deleted all of our video files, the analysis pipeline, and EthoVision output files, which all needed to be recovered from backup versions (if available) and reinstalled onto the lab's computer. This rebuilding of our analytical pipeline took some time to recover from in early Spring 2020. Moreover, the lab's catalog of videos (which have large file sizes) and the proprietary EthoVision software itself

have been unavailable as of March 7th, due to the COVID-19 pandemic and subsequently widely-enforced social distancing measures.

These setbacks have prohibited me from completing the analysis of the *entire* set of over 750 animals. The results I have presented are therefore preliminary, but readily scalable had circumstances allowed. The current best method for setting up EthoVision would be to use either the 0.05 cm/s movement threshold with 0.05% activity threshold (Experiment 4 parameter set 1) or the 0.07 cm/s movement threshold with 0.15% activity threshold (Experiment 5 parameter set 3). This parameter set is not perfect, however, as seen by the biases that its OLP regression contains (**Figure 12** A, C, and E). This indicates that further refinement of the movement and activity threshold needs to be performed in order to confidently choose a combination that works best to accurately call all types of behavior (but especially grooming). Additionally, further statistical analyses should be performed on those future data sets.

Finally, I recommend a subsequent sensitivity analysis, similar in scope to the one presented in this project, with finer grades of increments between settings. For example, a variation of movement thresholds from 0.03 – 0.09 cm/s with 0.01 cm/s steps to verify the proper setting for that initial classification of walking behavior with the expanded data set, and subsequent variation of the activity thresholds from 0.01 – 0.20% in 0.02% increments to refine the calls of grooming and standing behaviors.

Discussion

The utility of EthoVision as a tool for accurately detecting grooming behaviors

Our results suggest that the accurate and automated extraction of spontaneous behaviors, including grooming, using EthoVision is possible, although the premature conclusion of the experiment limits this interpretation, and some parameters have yet to be fully optimized with

respect to the entire available data set. The indices that EthoVision calculates for individual flies are in some cases very close to the manually called indices, and generally fall within statistical bounds on a population level with our initial training set (**Figure 10**). We concluded that our initial logic for determining behavior states using three different metrics in EthoVision's data profiles was far too restrictive. As a consequence, not every second of the videos were being put into a result, leading to total times called per arena of around only 500 seconds in some cases, which is only around 80% of the total length of the videos. After reevaluating our logic and discussing our results with representatives of Noldus Inc., while attending the Society for Neuroscience conference in October 2019, we decided to alter our logic by eliminating the mobility state variable and activity requirement for walking (**Figure 5**). This new logic greatly improved the number and accuracy of called behaviors, with on average only ~3 seconds (or <1%) of dropped time per fly, which in our assessment is an acceptable and unavoidable margin of error (Table 2B).

The rationale for our simplified logic is as follows: if the fly is moving, it is walking; if the fly is not moving, it is either inactive and classified as standing or has active appendages and is therefore grooming. From an analysis of several flies during grooming behaviors, we observed that high activity values indicated that the legs were moving, which is indicative of grooming. If there is a very low activity state value, then the fly is quite still and classified as standing. This new logic, in combination with a lower movement threshold value, similar to one that has been previously validated (0.07 cm/s; McLaughlin 2018), has led to better walking indices being called by EthoVision than in initial trials (**Table 3**). **Table 3** shows the indices resulting from the movement threshold of 0.05 cm/s. The absolute differences in all of the behavior indices are much smaller than those from the parameter set containing a movement threshold of 0.11 cm/s,

shown in **Table 4**. Future studies should vary the movement threshold around the 0.07 cm/s setting in smaller increments to determine the true best setting for determining walking in the larger data set.

The benefits of our approach

There are many benefits that come with our approach as compared to other automated analyses. First, our approach is accessible to an undergraduate laboratory as it is relatively user-friendly (only requiring manipulation of EthoVision's graphical user interface). Second, our method does not require high computational power; it can be executed on a standard PC. Third, our method allows for the use of standard and available recording equipment and arena setup that do not require expensive or specific video cameras. Finally, like the previously published automated approaches described above (Ramazani et al. 2007; Kain et al. 2013; Qiao et al. 2018), our automated analysis offers a high throughput and repeatable outcomes. Additionally, it is scalable to allow for analysis of large numbers of flies.

Problems with our approach

The main issue with our approach is that the initial setup in EthoVision can be tedious and is user-dependent, which allows for the possibility of human error. In the future, when an experiment is run with a new batch of videos from newly recorded flies, these videos would need to be manually loaded into the EthoVision software. The parameters for detection of behaviors would be the same, once they are thoroughly optimized, but the initial detection of the animals in their arenas depends on the person who inputs the videos because of minor variations in the brightness and focal range of individual recordings. These detection values will be similar from video to video and experiment to experiment, though they are likely to vary slightly depending on the specific user. This likely will not be a significant issue as manual scoring would lead to

the same if not greater discrepancies. The effects of these initial settings on subsequent analysis of behaviors, however, should be determined empirically in the future to ensure that differences in initial software settings do not lead to incorrect interpretations of EthoVision output.

The utility of accurately interpreting behaviors in fly research

Using *Drosophila melanogaster* in translational medicine has many advantages. Several fundamental cellular and developmental pathways are conserved in mammals and *Drosophila,* and many genes known to cause disease in humans have an orthologous or related gene in flies (Pandey & Nichols 2011). This is particularly true for genes involved in nervous system function and those that lead to intellectual disabilities and other cognitive disorders in humans (Inlow & Restifo 2004; van der Voet *et al.* 2014). Additionally, *Drosophila* reproduces rapidly and has several genetic systems in place that allow for easy manipulation of gene expression and phenotypic screening, such as RNAi libraries and the GAL4/UAS system (Pandey & Nichols 2011). This experimental control allows for the discovery and testing of drug treatments for various diseases. McBride *et al.* (2005) provide one such example, where the authors found several potential drug treatments for fragile X syndrome based on amelioration of cognitive deficits in a learning and memory assay, some of which have made it to the clinical trial phase. Several promising treatments were found because they targeted the grooming phenotype, which is evidence that studying behavior in fly models is a worthwhile endeavor (Tauber *et al.* 2011; McBride *et al.* 2013).

Furthermore, *Drosophila* as a model organism can be used for both the initial high throughput screen for drug candidates and the secondary validation of biological compounds (Pandey & Nichols 2011). One such example of the latter was screening of several compounds in a fly model of Huntingdon's Disease. If the compounds were not effective or were toxic, they

would not have progressed onto more resource-intensive rodent models or further clinical trials in humans (Pandey & Nichols 2011).

Future implementation of this approach and extensions

The analysis pipeline outlined in this work is readily available for implementation on the larger video data set available in the Andrew lab. With minimal training in the software and analysis pipeline, the documentation I have provided should allow future students to continue and expand upon this work. Future work includes refining the threshold values in order to more accurately call standing and grooming bouts. This might include incorporating falling into the EthoVision called behaviors. The falling index is currently not accounted for in EthoVision, but it was called during the manually scored videos because some flies attempt to walk along the glass coverslip surface of the arena and fall onto the substrate (**Table 5**). This was observed as a common occurrence for several seemingly uncoordinated lines of the DGRP in the early analysis of grooming from a prior student's thesis (Hannum 2017). Ultimately, the lab is currently focused on the extraction of grooming indices, but the incorporation of falling into EthoVision for the sake of current validation might be necessary until it is determined that the lack of falling index is not disrupting the grooming indices.

Additionally, there are times when the fly exhibits no movement coupled with high activity when it is not grooming. This could occur in instances where the fly simply moves its abdomen or wings which results in a pixel change large enough to fall above the activity threshold set for grooming. Furthermore, a fly that has fallen onto its back and is moving its legs a lot in an effort to right itself might also appear to be grooming based on the logic set forth for the grooming behavior. Moving forward, this issue will need to be addressed in order to not only

more accurately call the grooming index, but also ensure that the grooming that EthoVision is calling consists of actual, true grooming bouts.

Acknowledgements

I would like to thank Dr. Andrew for allowing me to take on this project and for all of his support and guidance throughout this process. I would also like to thank Dr. Brandon, Dr. McDonald, and Dr. Morrison for being members of my Honors Committee. A special thanks to Sohini Mukherjee and Jose Martinez for caring for the lab's fly stocks and to Courtney Hannum and Sean McLaughlin for their previous work that led up to this project. Finally, thank you to the Lycoming College Biology Department and the Haberbergers for providing funding for this project.

Figures and Legends

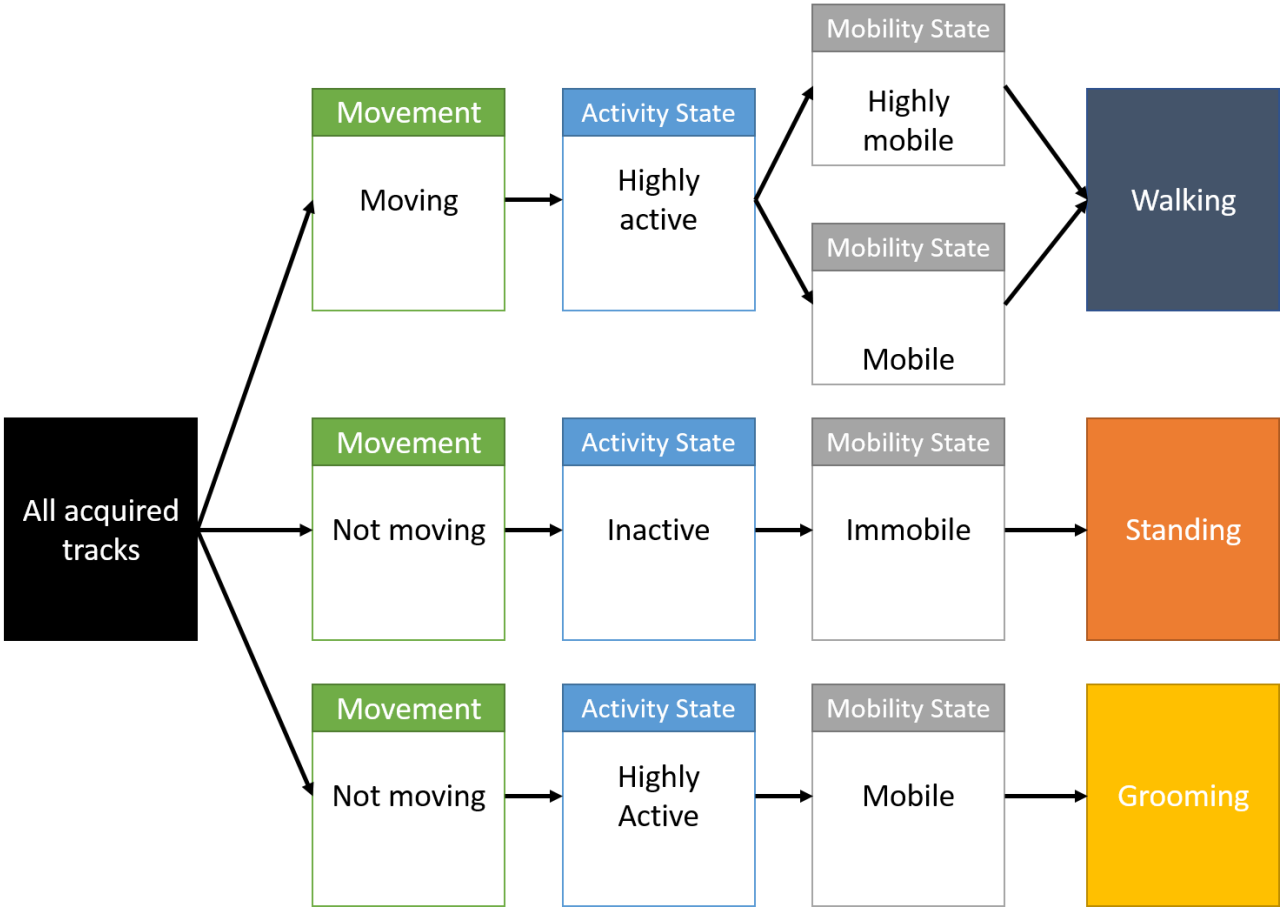


Figure 2. Flow diagram illustrating the initial logic for determining behavior. The frames in each video are compared to threshold values, which determine what "result" a certain frame should fall into. Grooming requires a "mobile" mobility state because the fly's body moves slightly while it is grooming. This pixel difference between two frames is enough to make the percent mobility above the immobile threshold.

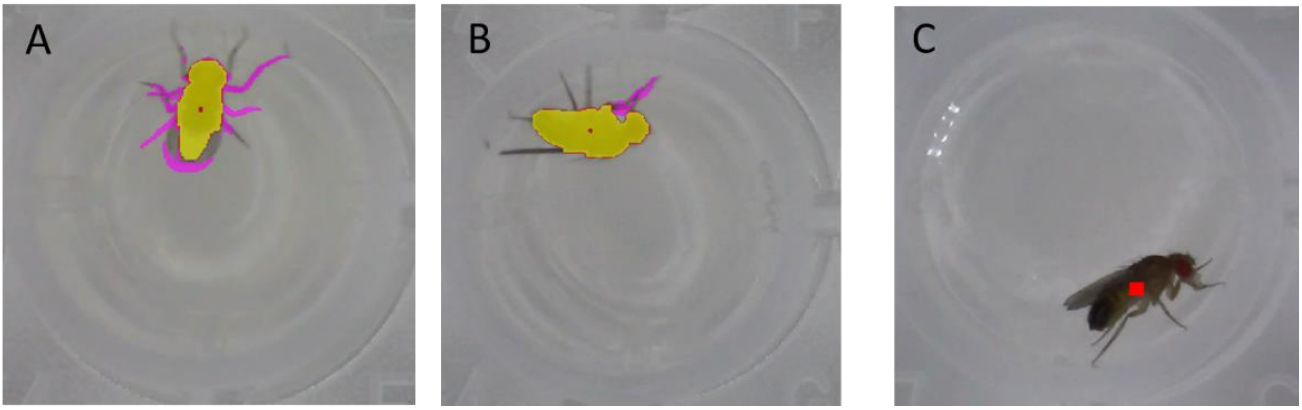


Figure 3. Representation of EthoVision's activity state (A & B) and movement (C) variables. Activity state is calculated by comparing the pixels in the entire arena from one frame to the next. **A)** The pixels in yellow are detected as the animal while the pixels in purple are the pixels in the arena that changed from the last frame. **B)** The pixels that changed during a grooming bout. In both cases (**A** and **B**) notice the detected activity during a walking bout (**A**) and grooming bout (**B**). **C)** An animal is moving if its average velocity exceeds the start velocity defined by the user. The red box on the animal is its center-point, which is used to track distance moved.

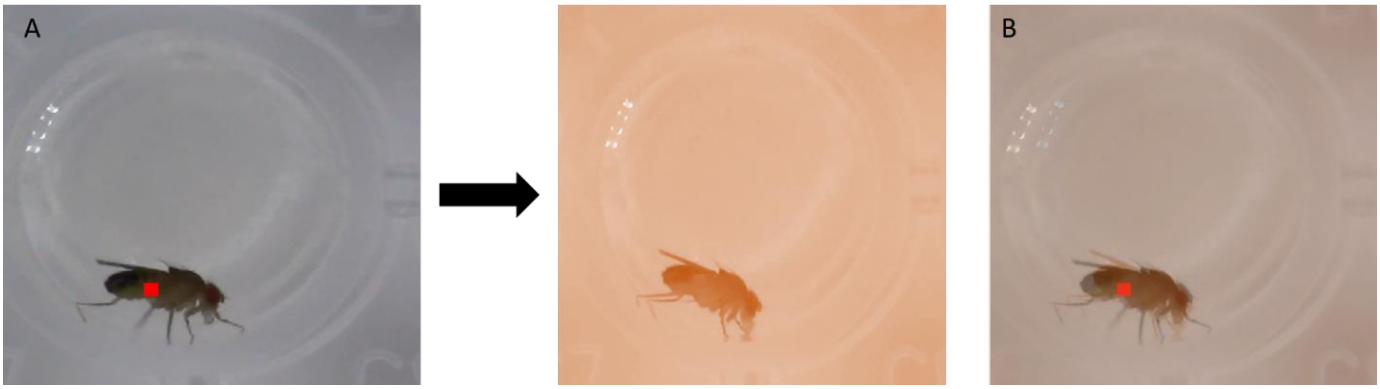


Figure 4. Representation of EthoVision's mobility state variable calculated by comparing the pixels defined as the subject in one frame to the subject's pixels in the next frame. The difference between the pixel values is used to calculate mobility state, which ranges from 0 (no difference in pixels) to 100 (every pixel moved). **A)** The fly in one frame compared to the same fly in the next frame of the video. **B)** An overlay of the two frames. Note the difference in the position of the fly's body. EthoVision uses this to calculate percent mobility. The user-defined threshold values are then used to calculate mobility state.

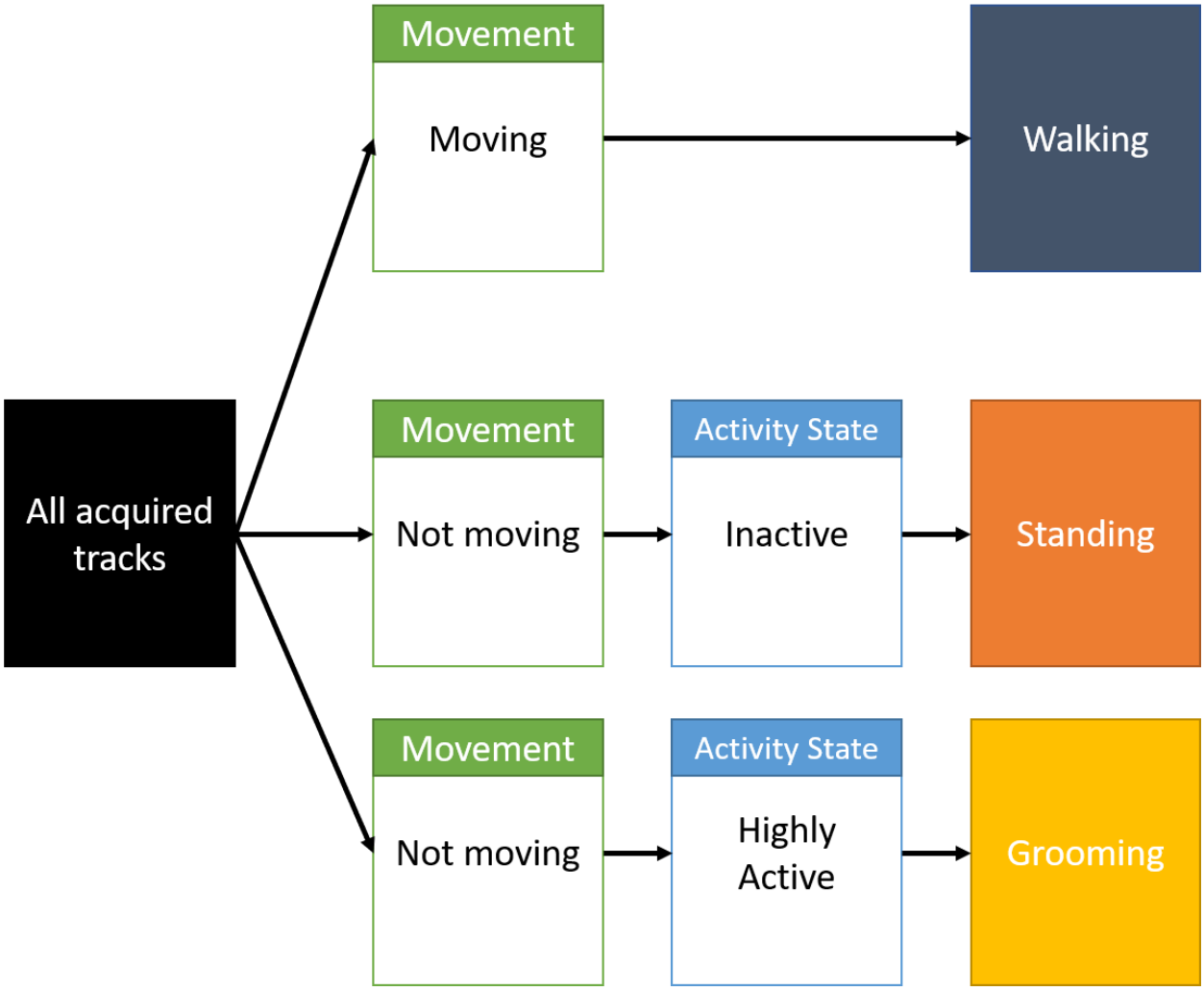


Figure 5. New logic for determining behavior. The previous logic was determined to be too restrictive. This is likely the cause of there being total times called per arena of less than the 600 second video length. This is due to the fact that it is much harder to meet three requirements than it is one. To simplify this logic and make the binning less restrictive, the mobility states were removed from all behaviors and the activity state was removed from walking.

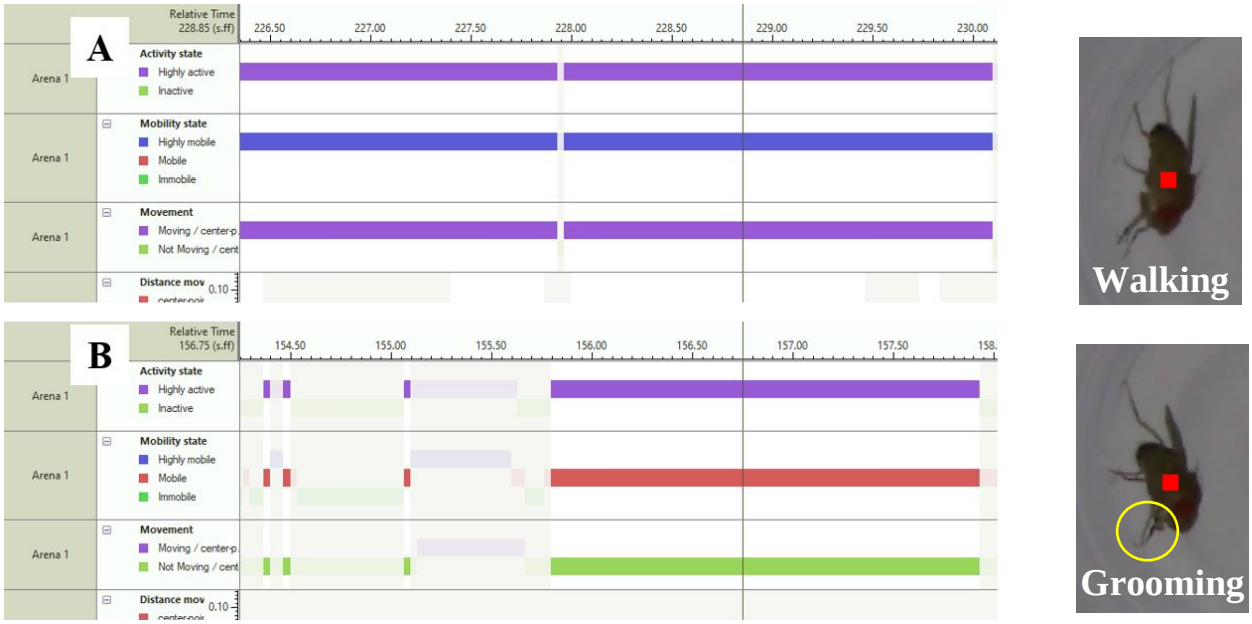


Figure 6. Track visualization in EthoVision. The lighter, not tan areas show which frames are called as that behavior. **A)** Activity and mobility states and movement shown for a fly (right) that is walking around the arena. **B)** Activity and mobility states and movement shown for a fly (right) during a grooming bout. Note the position of the fly's front legs, circled in yellow.

	Experiment 1: Activity	**Experiment 2: Mobility**	**Experiment 3: Mobility**
Movement thresholds	Start: 0.11cm/s Stop: 0.11cm/s	Start: 0.11cm/s Stop: 0.11cm/s	Start: 0.11cm/s Stop: 0.11cm/s
Activity thresholds	Inactive below varied from 0.05% to 0.65% in increments of 0.1%	Inactive below 0.05%	Inactive below 0.05%
Mobility thresholds	High above 5% Immobile before 1%	High above varied from 5% to 12% in increments of 1% Immobile below 1%	High above 5% Immobile below varied from 1% to 0.3% in increments of 0.1%

Table 1. We compared the results for grooming, standing, and walking that EthoVision provided at different parameter settings to manually-scored results. This table shows thresholds defining each of the behaviors that EthoVision calls. We held two thresholds constant in each experiment, while varying only one of them (shown in red). For example, in Experiment 1, the movement threshold was 0.11 cm/s and the mobility thresholds were 5% and 1% for every parameter set, while the activity threshold was varied from 0.05% to 0.65% in increments of 0.1%.

A

Trial	Arena	Total time (s)
1	1	576.2945
	2	580.0943
2	1	525.928
	2	555.294
3	1	520.3283
	2	544.5277
4	1	568.0605
	2	501.1619
5	1	570.56105
	2	575.62714
	3	576.0944
6	1	549.3613
	2	531.1949
Min value (s)		**501.1619**
Max value (s)		**580.0943**
Average (s)		**551.8868**
Standard dev.		**24.6427**

B

Trial	Arena	Total time (s)
1	1	598.4609
	2	598.8603
2	1	596.7937
	2	598.1271
3	1	596.1612
	2	597.2941
4	1	598.561
	2	595.6939
5	1	599.1278
	2	599.0942
	3	599.4611
6	1	597.594
	2	597.794
Min value (s)		**595.6939**
Max value (s)		**599.4611**
Average (s)		**597.9249**
Standard dev.		**1.1788**

Table 2. The total time called by EthoVision for a selection of arenas. **A)** These times were calculated using the initial EthoVision logic. Note how few seconds were actually called. Some arenas contained fewer than 550 seconds, with the lowest being 501 seconds. The average represents just 92% of the total time present in the video (551 seconds compared to the total 600 seconds). The standard deviation is also high which demonstrates that the total time called by EthoVision varied throughout the experiment. **B)** These times were calculated using the simplified EthoVision logic. Note that these times were much closer to the total 600 seconds, with the average being 597 seconds.

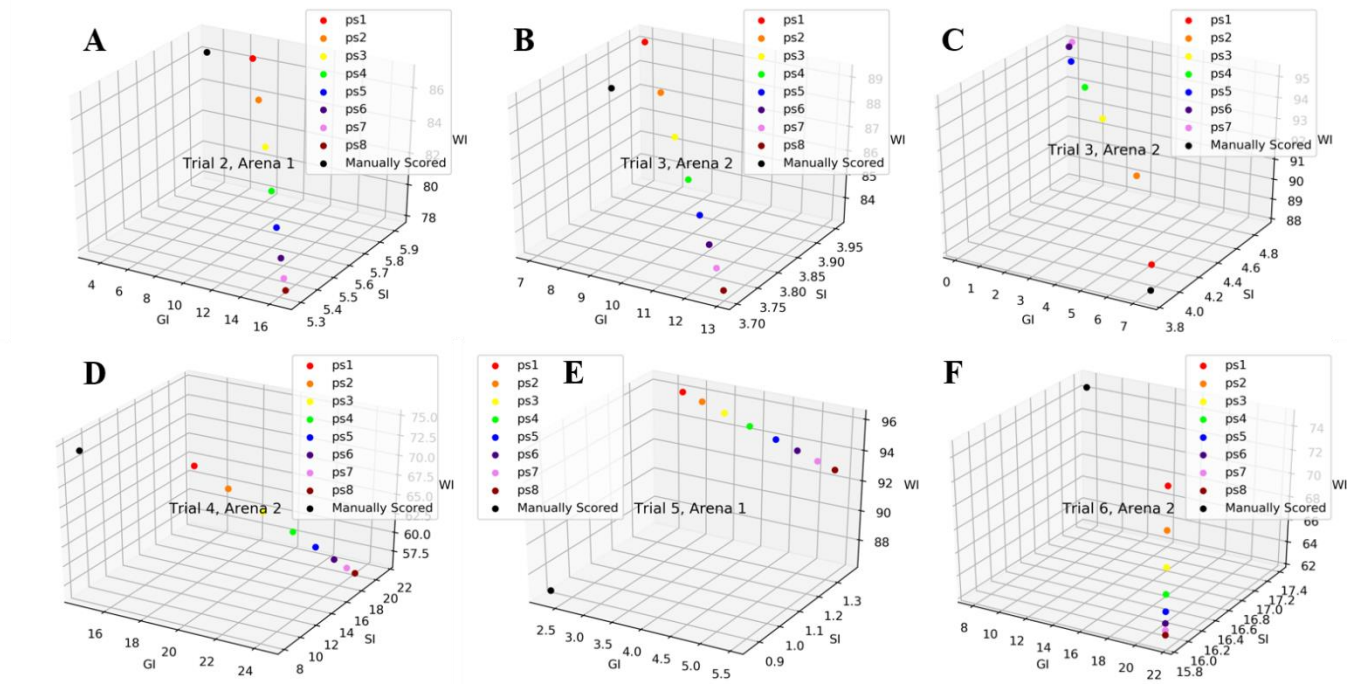


Figure 7. Three-dimensional scatter plots showing a comparison of the parameter sets for certain flies, taken from Experiment 2: mobility (A, B, and D-F) and Experiment 1: Activity (C). Each plot represents the data for a different fly, corresponding to the trial and arena where the fly is found. Each of the behavioral indices are on each of the axes (grooming, standing, and walking). The dots represent the combination of these three indices in 3D space. **A-C)** These parameter sets (colors) are close to the manually called value (black). **D-F)** These parameter sets are noticeably different from the manually called value.

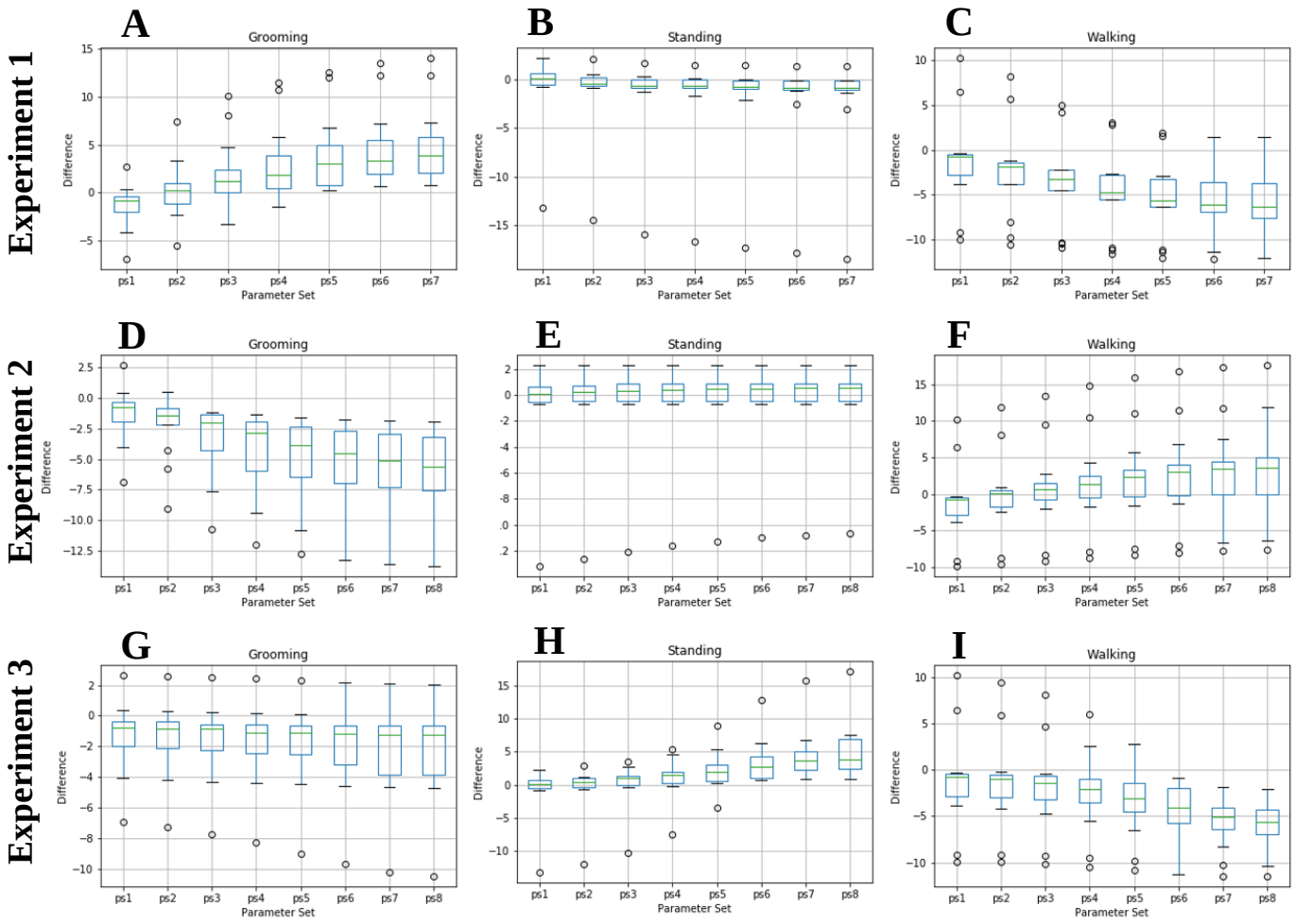


Figure 8. We compared the measures for grooming, standing, and walking that EthoVision provided at different threshold values (**Table 1**) to manually-scored measures (n = 13). This figure features box plots showing the difference between the behavior index calculated using the time called by EthoVision and the manually called indices for each behavior. Experiment 1 (from **Table 1**) is illustrated along the top (A-C), experiment 2 is along the middle (D-F), and experiment 3 is shown along the bottom (G-I).

Behavior	Trial	Arena	Index	Manually scored	Difference	% Difference
Grooming	2	1	4.8336	3.34	-1.4936	44.719
		2	6.58925	5.75	-0.83925	14.596
	3	1	8.46017	12.33	3.86983	31.385
		2	5.82713	7.37	1.54287	20.934
	10	1	5.52808	5.94	0.41192	6.935
		2	4.93916	6.19	1.25084	20.207
	14	1	12.2396	15.38	3.1404	20.419
		2	12.4229	11.77	-0.6529	5.547
		3	10.5061	13.73	3.2239	23.481
Standing	2	1	5.66698	5.93	0.26302	4.435
		2	3.07795	3.74	0.66205	17.702
	3	1	5.69937	5.44	-0.25937	4.768
		2	3.88846	3.85	-0.03846	0.999
	10	1	2.07234	4.15	2.07766	50.064
		2	1.77232	1.52	-0.25232	16.600
	14	1	12.4896	16.82	4.3304	25.746
		2	80.6545	84.49	3.8355	4.540
		3	5.90588	7.29	1.38412	18.987
Walking	2	1	88.9772	86.09	-2.8872	3.354
		2	90.0328	87.41	-2.6228	3.001
	3	1	85.2961	81.16	-4.1361	5.096
		2	89.8122	88.29	-1.5222	1.724
	10	1	92.1051	88.65	-3.4551	3.897
		2	93.1052	88.39	-4.7152	5.335
	14	1	74.8597	66.01	-8.8497	13.407
		2	6.00589	3.61	-2.39589	66.368
		3	83.1991	76.96	-6.2391	8.107

Table 3. Parameter set 1 shown for 9 flies. The movement variable was averaged over 3 samples and featured a start velocity and stop velocity both at 0.05cm/s. Activity state was averaged over 3 samples with inactivity below 0.05% and excluding instances shorter 0.10 sec. The difference was calculated by subtracting the index from the manually scored data. The percent difference was calculated by subtracting the manually scored from the index and dividing by the manually scored. The smaller average percent difference (compared to **Table 4**) is indicative of a better fit of the detection settings.

Behavior	Trial	Arena	Index	Manually scored	Difference	% Difference
Grooming	2	1	16.762	3.34	-13.422	401.856
		2	13.4619	5.75	-7.7119	134.120
	3	1	18.4146	12.33	-6.0846	49.348
		2	13.443	7.37	-6.073	82.402
	10	1	10.9562	5.94	-5.0162	84.448
		2	11.9618	6.19	-5.7718	93.244
	14	1	23.8235	15.38	-8.4435	54.899
		2	15.2064	11.77	-3.4364	29.196
		3	21.879	13.73	-8.149	59.352
Standing	2	1	6.02256	5.93	-0.09256	1.561
		2	3.2224	3.74	0.5176	13.840
	3	1	6.17154	5.44	-0.73154	13.447
		2	4.29952	3.85	-0.44952	11.676
	10	1	2.27235	4.15	1.87765	45.245
		2	1.989	1.52	-0.469	30.855
	14	1	12.6563	16.82	4.1637	24.754
		2	80.7712	84.49	3.7188	4.401
		3	6.15034	7.29	1.13966	15.633
Walking	2	1	76.6043	86.09	9.4857	11.018
		2	82.9768	87.41	4.4332	5.072
	3	1	74.5917	81.16	6.5683	8.093
		2	81.5021	88.29	6.7879	7.688
	10	1	86.327	88.65	2.323	2.620
		2	85.7548	88.39	2.6352	2.981
	14	1	63.0591	66.01	2.9509	4.470
		2	3.04461	3.61	0.56539	15.662
		3	71.4484	76.96	5.5116	7.162

Table 4. Parameter set 7 shown for 9 flies. The movement variable was averaged over 3 samples and featured a start velocity and stop velocity both at 0.11cm/s. Activity state was averaged over 3 samples with inactivity below 0.05% and excluding instances shorter 0.10 sec. The difference was calculated by subtracting the index from the manually scored data. The percent difference was calculated by subtracting the manually scored from the index and dividing by the manually scored. The very large percent differences compared to the results in **Table 3**, particularly in regard to grooming, indicate that this parameter set inaccurately detects behaviors.

Behavior	Trial	Arena	Index
Falling	2	1	4.62
		2	3.09
	3	1	1.08
		2	0.45
	10	1	1.26
		2	3.89
	14	1	1.79
		2	0.12
		3	2.02

Table 5. Manually scored falling index from MF (Trials 2 and 3) and SM (Trials 10 and 14) (McLaughlin 2018). Due to it being difficult to classify falling given EthoVision's parameters, the falling index was omitted from the EthoVision analysis. Due to some of the fly's having a substantial falling index, for instance those that are above 1, the omission of this behavioral index might be leading to some of the discrepancies seen with the other behavioral indices.

	Experiment 4: 0.05cm/s	**Experiment 5: 0.07cm/s**
Movement thresholds	Start: 0.05cm/s Stop: 0.05cm/s	Start: 0.07cm/s Stop: 0.07cm/s
Activity thresholds	Inactive below varied from 0.05% to 0.45% in increments of 0.05%	Inactive below varied from 0.05% to 0.45% in increments of 0.05%

Table 6. We compared the results for grooming, standing, and walking that EthoVision provided at different parameter settings to manually-scored results. This table was used with the new EthoVision logic and shows thresholds defining each of the behaviors that EthoVision calls. We held one of the thresholds constant in each experiment, while varying the other one (shown in red). For example, in experiment 4, the movement threshold for each parameter set was 0.05 cm/s while the activity threshold varied from 0.05% to 0.45% in increments of 0.05%. So parameter set 1's activity threshold would have been 0.05%, parameter set 2's 0.10%, parameter set 3's 0.15%, and so on.

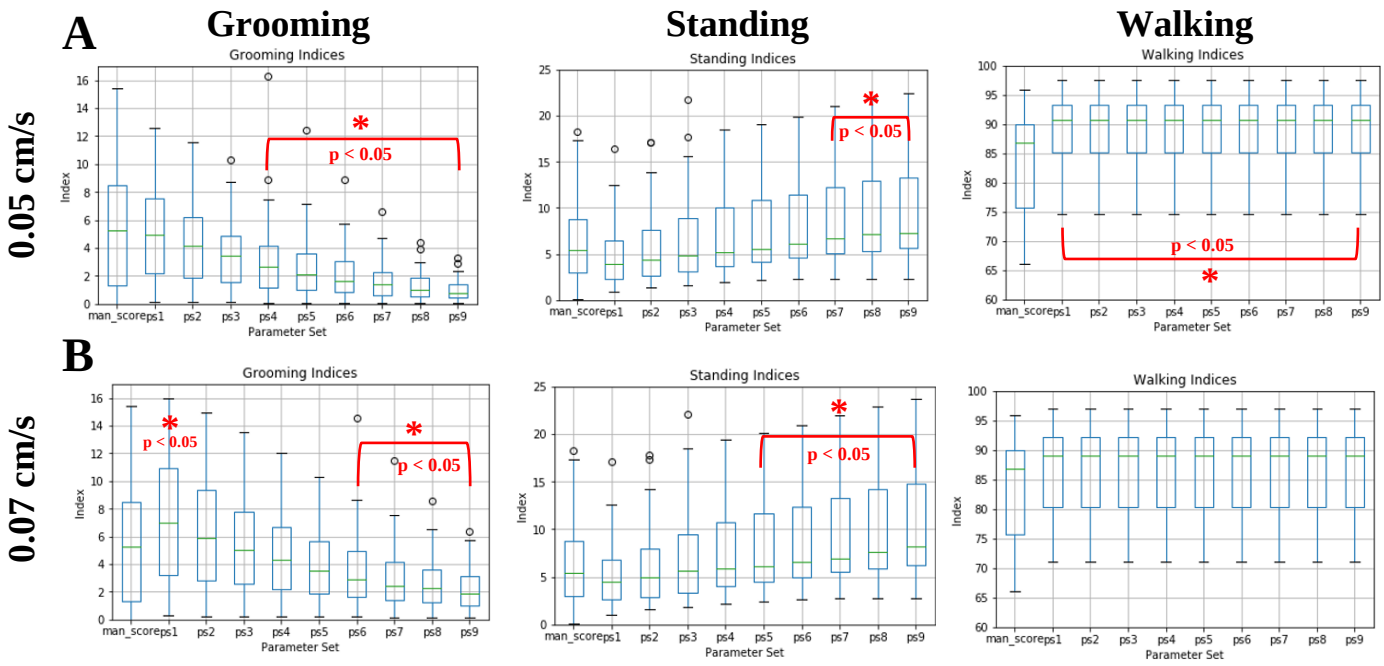
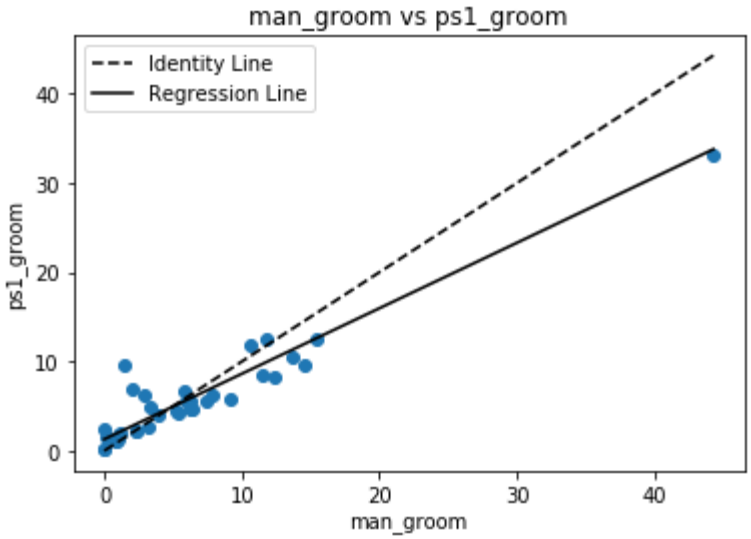


Figure 9. Box plots comparing the manual scores and the EthoVision scores for grooming, standing, and walking, respectively (n = 35). Within each box plot, the manual scores are shown first, with the 9 EthoVision scored parameter sets shown afterward. For each box on the plot, the spread of scores are shown. If the EthoVision parameters are calling behaviors accurately, their spread should look similar to the manual score's spread (first box). The EthoVision parameter sets for walking are all the same within each experiment because the activity threshold is varied, meaning only the results affected by the activity variable will change. A red asterisk indicates statistically significant pairwise differences ($p < 0.05$) between the manual scores and each parameter set. **A)** Experiment 4, with movement threshold of 0.05cm/s. Note that each parameter set for walking is the same, meaning they are all statistically different than the manual scores. **B)** Experiment 5, with movement threshold of 0.07cm/s.

A

```
a:  1.2912446310985155
b:  0.7337323261466941
a_CI95:  [0.7158554071236551, 1.8043544752529437]
b_CI95:  [0.6543140952821253, 0.8227900488686688]
```



B

```
pyc.blandAltman(man_groom, ps1_groom)
```

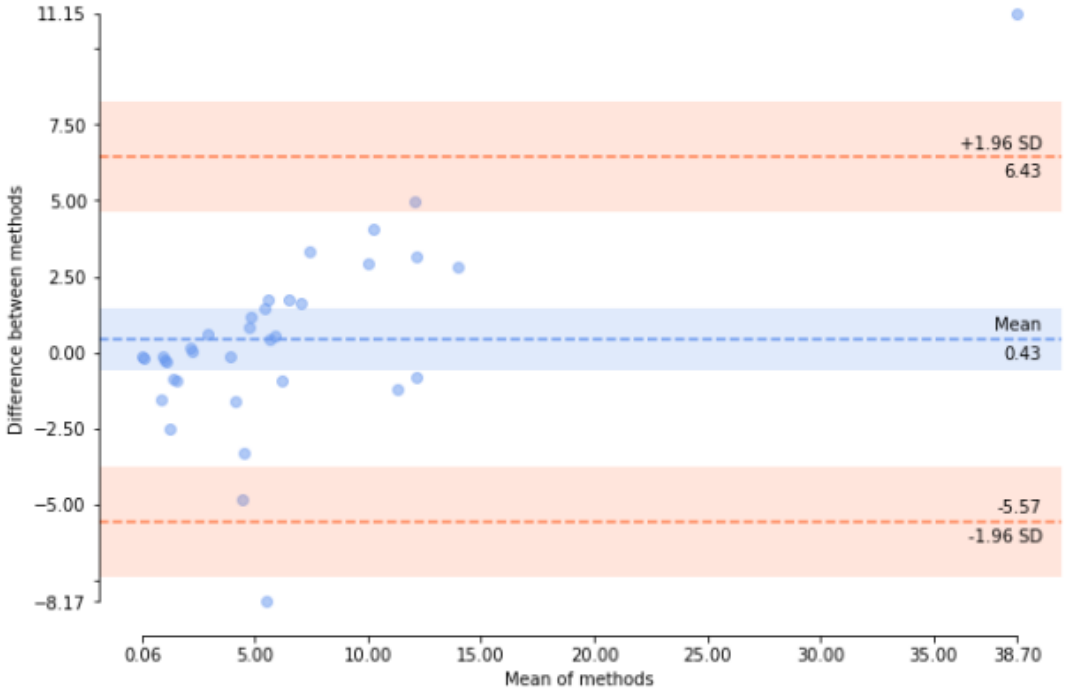
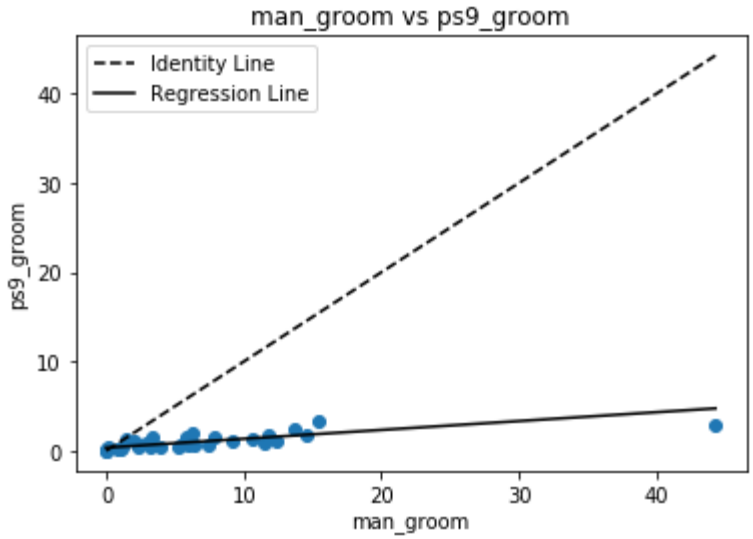


Figure 10. Results comparing the manual scores and the EthoVision scores for grooming with the first parameter set in Experiment 4, where the movement threshold was 0.05cm/s. The activity threshold for this experiment was 0.05%. **A)** Ordinary least products regression analysis performed on the same datasets as in A. The identity line (y = x) is the dotted line while the regression line is the solid line. The dots are the data points. **B)** Bland Altman plot showing the mean of the scores on the x axis and the difference between the manual scores and the EthoVision scores on the y axis.

A

```
a:  0.35734553998463814
b:  0.09922453942660268
a_CI95:  [0.1865266066155512, 0.4922249421440782]
b_CI95:  [0.07834814498278539, 0.12566359071532174]
```



B

```
pyc.blandAltman(man_groom, ps9_groom)
```

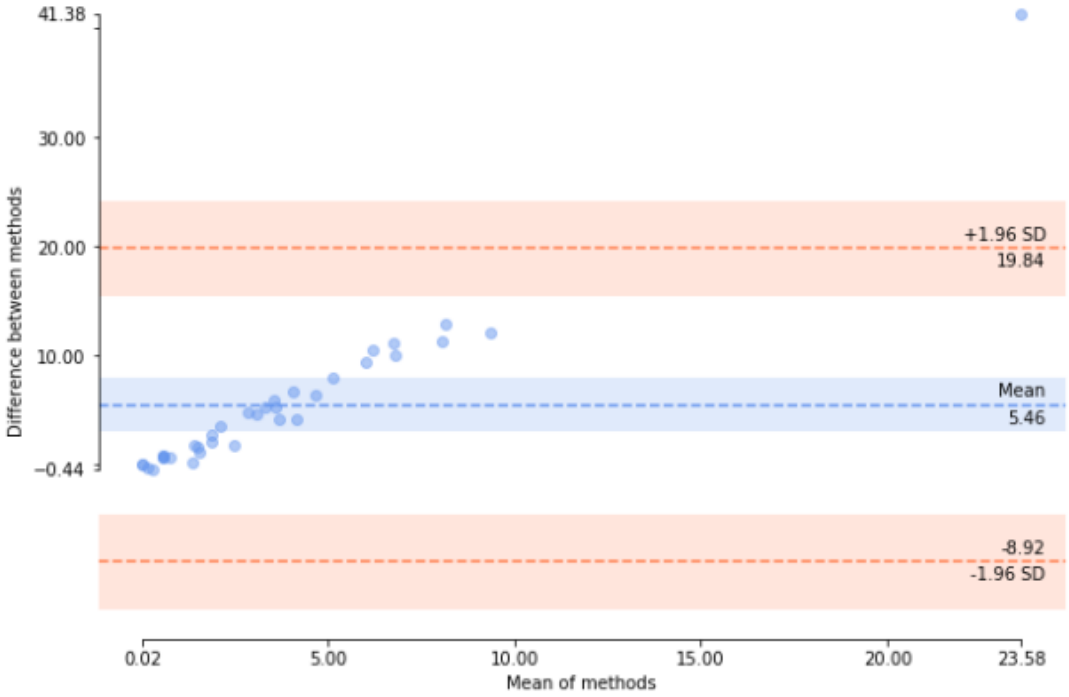


Figure 11. Results comparing the manual scores and the EthoVision scores for grooming with the final parameter set in Experiment 4, where the movement threshold was 0.05cm/s. The activity threshold for this experiment was 0.45%. **A)** Ordinary least products regression analysis performed on the same datasets as in A. The identity line (y = x) is the dotted line while the regression line is the solid line. The dots are the data points. **B)** Bland Altman plot showing the mean of the scores on the x axis and the difference between the manual scores and the EthoVision scores on the y axis.

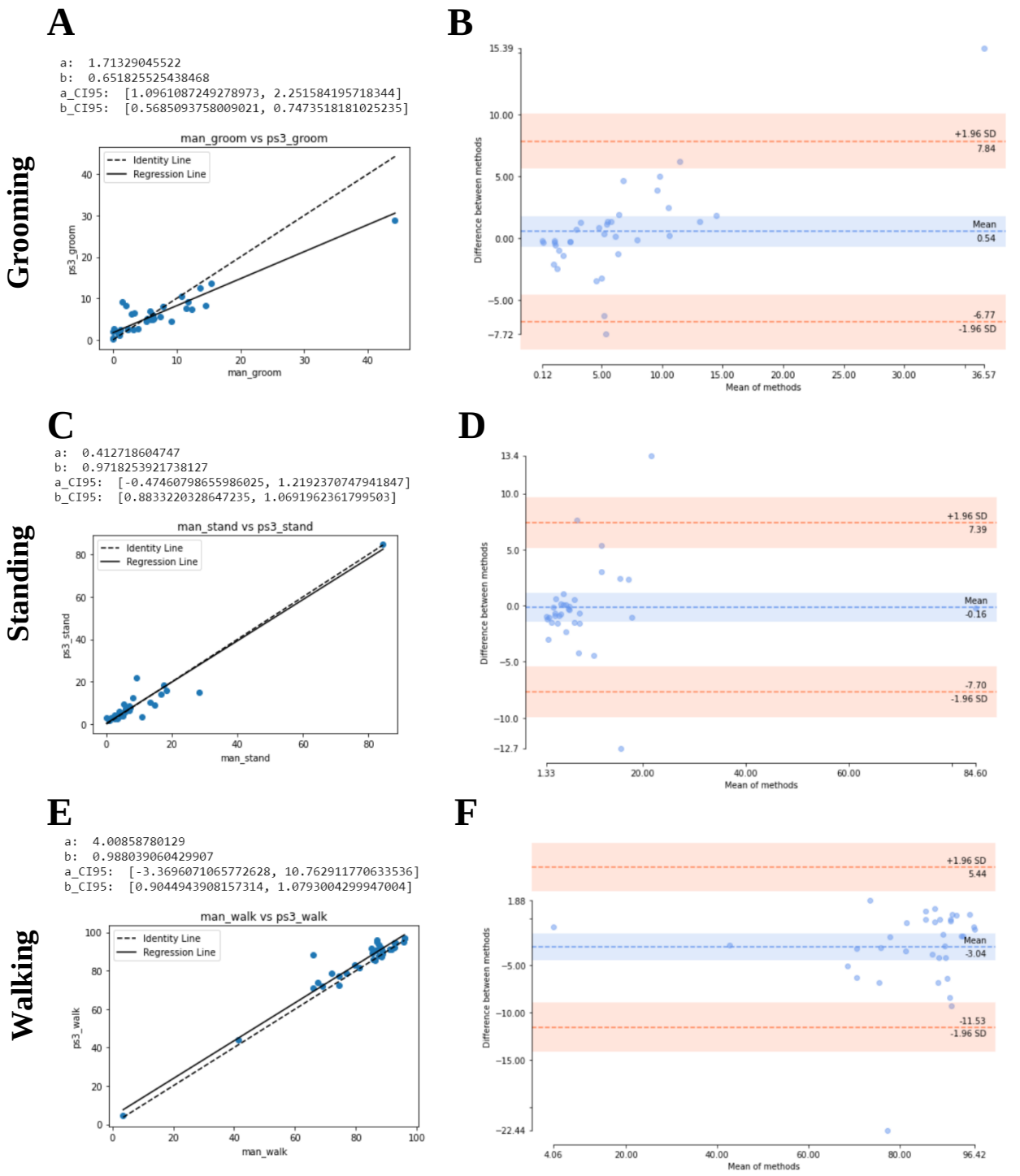


Figure 12. Results comparing the manual scores and the EthoVision scores for all three behaviors for parameter set 3 in Experiment 5, where the movement threshold was 0.07cm/s and the activity threshold was 0.15%. **A** and **B** show grooming, **C** and **D** show standing, and **E** and **F** show walking. The first column (**A**, **C**, and **E**) are ordinary least product regression analyses and the second column (**B**, **D**, and **F**) show Bland-Altman plots.

References

Altman, D.G. and J.M. Bland. (1983) Measurement in Medicine: the Analysis of Method Comparison Studies. The Statistician. 32: 307-317. https://www.jstor.org/stable/2987937

American Psychiatric Association. (2013) Diagnostic and Statistical Manual of Mental Disorders (5th ed.). Washington, D.C.: American Psychiatric Association.

Andrew, D.R., M.E. Moe, D. Chen, J.A. Tello, R.L. Doser, *et al.* Spontaneous motor-behavior abnormalities in two *Drosophila* models of brain-development disorders. Preparing for submission to the Journal of Neurogenetics.

Dawkins, R. and M. Dawkins. (1976) Hierarchical organization and postural facilitation: rules for grooming in flies. Anim. Behav. 24: 739-755.

Garber, K.B., J. Visootsak, and S.T. Warren. (2008) Fragile X Syndrome. Eur. J. Hum. Genet. 16: 666-672. doi: 10.1038/ ejhg.2008.61

Giavarina, D. (2015) Understanding Bland Altman analysis. Biochemia Medica. 23(2): 141-51. http://dx.doi.org/10.11613/BM.2015.015

Hagedorn, J., J. Hailpern, and K.G. Karahalios. (2008) VCode and VData: illustrating a new framework for supporting the video annotation workflow. Extended Abstracts of AVI.

Hales, K., C. Korey, A. Larracuente, D. Roberts. (2015) Genetics on the Fly: A Primer on the *Drosophila* Model System. Genetics. 201(3): 815-842. https://doi.org/10.1534/genetics.115.183392

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

Huang, W., A. Massouras, Y. Inoue, J. Peiffer, M. Ràmia, *et al.* (2014) Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. Genome Res. 24(7): 1193-1208. doi: 10.1101/gr.171546.113

Inlow, J.K. and L.L. Restifo. (2004). Molecular and comparative genetics of mental retardation. Genetics. 166(2): 835-881. doi: 10.1534/genetics.166.2.835

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). doi: 10.1038/ncomms2908

Ludbrook, J. (1997) Comparing Methods of Measurement. Clin. Exp. Pharmacol. P. 24: 193-203. https://doi.org/10.1111/j.1440-1681.1997.tb01807.x

Mackay, T. F. C. (2009) The genetic architecture of complex behaviors: lessons from *Drosophila*. Genetica. 136: 295-302. doi: 10.1007/s10709-008-9310-6

Mackay, T. F. C. and R. R. H. Anholt. (2006) Of flies and man: *Drosophila* as a model for human complex traits. Annu. Rev. Genomics Hum. Genet. 7: 339-67. doi: 10.1146/annurev.genom.7.080505.115758

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. doi:10.1038/nature10811

Mann, H.B. and D.R. Whitney. (1947) On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics. 18(1): 50-60. https://www.jstor.org/stable/2236101

McBride, S.M.J., C.H. Choi, Y. Wang, D. Liebelt, E. Braunstein, *et al.* (2005) Pharmacological rescue of synaptic plasticity, courtship behavior, and mushroom body defects in a *Drosophila* model of Fragile X Syndrome. Neuron. 45: 753-764. doi: 10.1016/j.neuron.2005.01.038

McBride, S.M.J., S.L. Holloway, and T.A. Jongens. (2013) Using *Drosophila* as a tool to identify pharmacological therapies for fragile X syndrome. Drug Discov. Today 10(1): e129-e136. doi: 10.1016/j.ddtec.2012.09.005

McLaughlin, S. P. (2018) Genetic analysis of spontaneous walking behavior in *Drosophila melanogaster*. Lycoming College Departmental Honors Thesis.

McKinney, W. (2010) Data Structures for Statistical Computing in Python. Proc. Of the 9th Python in Science Conf 51-56. doi: 10.25080/Majora-92bf1922-00a

Nachar, N. (2008) The Mann-Whitney U: a test for assessing whether two independent samples come from the same distribution. Tutor. Quant. Methods. Psychol. 4(1): 13-20. http://dx.doi.org/10.20982/tqmp.04.1.p013

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. doi: 10.1124/pr.110.003293

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. 7: e34497. https://doi.org/10.7554/eLife.34497

Ramazani, R.B., H.R. Krishnan, S.E. Bergeson, and N.S. Atkinson. (2007) Computer automated movement detection for the analysis of behavior. J. Neurosci. Meth. 162: 171-179. doi:10.1016/j.jneumeth.2007.01.005

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. http://dx.doi.org/10.7554/eLife.02951

Szebenyi, A.L. (1969) Cleaning behaviour in *Drosophila melanogaster*. Anim. Behav. 17: 641-651.

Tauber, J.M., P.A. Vanlandingham, and B. Zhang. (2011) Elevated levels of the vesicular monoamine transporter and a novel repetitive behavior in the *Drosophila* model of fragile X syndrome. PLoS One 6(11): e27100. doi: 10.1371/journal.pone.0027100

Voet, M., B. Nijhof, M. Oortveld, and A. Schenck. (2014) *Drosophila* models of early onset cognitive disorders and their clinical applications. Neurosci. Biobehav. R. 46 pt. 2. 2: 326-342. https://dx.doi.org/10.1016/j.neubiorev.2014.01.013

URLs

Drosophila Genetics Reference Panel GWAS webtool: http://dgrp2.gnets.ncsu.edu

Pearce, J.T.M. (2019) pyCompare. https://pypi.org/project/pyCompare/

Trujillo-Ortiz, A. (2020) gmregress. https://www.mathworks.com/matlabcentral/fileexchange/27918-gmregress, MATLAB Central File Exchange.