Action detection using a neural network elucidates the genetics of mouse grooming behavior

Brian Q. Geuther¹, Asaf Peer¹, Hao He¹, Gautam Sabnis¹, Vivek M. Philip¹, Vivek Kumar¹*

¹The Jackson Laboratory, Bar Harbor, ME, USA

Abstract  Automated detection of complex animal behaviors remains a challenging problem in neuroscience, particularly for behaviors that consist of disparate sequential motions. Grooming, a prototypical stereotyped behavior, is often used as an endophenotype in psychiatric genetics. Using mouse grooming behavior as an example, we develop a general purpose neural network architecture capable of dynamic action detection at human observer-level performance and operate across dozens of mouse strains with high visual diversity. We provide insights into the amount of human annotated training data that are needed to achieve such performance. We survey grooming behavior in the open field in 2500 mice across 62 strains, determine its heritable components, conduct GWAS to outline its genetic architecture, and perform PheWAS to link human psychiatric traits through shared underlying genetics. Our general machine learning solution that automatically classifies complex behaviors in large datasets will facilitate systematic studies of mechanisms underlying these behaviors.

Introduction

Behavior, the primary output of the nervous system, is complex, hierarchical, dynamic, and high dimensional (Gomez-Marin et al., 2014). Precise approaches to dissect neuronal function requires analysis of behavior at high temporal and spatial resolution. Achieving this is a time-consuming task and its automation remains a challenging problem in behavioral neuroscience. In the field of computer vision, modern neural network approaches have presented new solutions to visual tasks that perform just as well as humans (Ching et al., 2018; Angermueller et al., 2016). Application of these tools to biologically relevant problems could alleviate the costs of behavioral experiments and enhance reproducibility. Despite these enticing advantages, few aspects of behavioral biology research leverages neural network approaches. This lack of application is often attributed to the high cost of organizing and annotating the data sets, or to the stringent performance requirements. Thus, behavior recognition within dynamic environments is an open challenge in the machine learning community and translatability of proposed solutions to behavioral neuroscience remains unaddressed.

Behavioral action recognition falls under multiple types of computer vision problems, including action classification, event detection, and temporal action localization. Action classification, a task closely related to image captioning, trains a classifier to apply action labels to manually pre-trimmed video clips. This problem has already been largely solved, with the exceptional performance for networks competing in data sets such as Kinetics-400, Moments in Time, Youtube-8M, and many other available benchmark data sets (Wu et al., 2017). However, this classification does not
determine when an action occurs within an untrimmed video. To address this shortcoming, two other tasks have been designed: event detection (ActivityNet 2019 Task 1) and temporal action localization (ActivityNet 2019 Task 2) (Fabricio Caba Heilbron and Niebles, 2015). The objective of event detection is to identify when an event occurs, while the objective of temporal action detection is to identify where, when, and who is performing an action in untrimmed video input. The dominant approach for solving these issues has been extending region proposal methods from single images to video data. This involves proposing video tubelets (Kalogeiton et al., 2017; Feichtenhofer et al., 2019), a clip of video in both space and time for a single subject performing a single action.

In behavioral neuroscience, previous attempts to operate directly on visual data have utilized unsupervised behavioral clustering approaches (Todd et al., 2017). These include seminal work to convert visual data into frequency domains followed by clustering in Drosophila (Berman et al., 2014) and autoregressive Hidden Markov Model-based analysis of depth imaging data for mouse behavior (Witschko et al., 2015). Both approaches rely upon alignment of data from a top-down view and while they cluster similar video segments, interpretation of generated clusters is still dictated by the user. It is also unclear how these approaches will perform on sequences of disparate behaviors.

Supervised approaches in behavioral neuroscience have abstracted the subject into lower dimensions such as ellipse or key points, followed by feature generation, and classification (Kabra et al., 2013; van den Boom et al., 2017). While these approaches were a significant advance when they were introduced, they are inherently limited by the measurements available from the abstraction. For instance, standard measurements such as center of mass tracking, limit the types of behaviors that can be classified reliably. The field quickly recognized this issue and moved to integrate new measurements for the algorithms to classify behavior. These new features are highly specific to the organism and behavior that the researcher wishes to observe. In Drosophila studies, tracking of individual limbs and wings add new tracking modalities (Robie et al., 2017). For mice, modern systems integrate floor vibration measurements and depth imaging techniques to enhance behavior detection (Quinn et al., 2003; Hong et al., 2015). Vibration measurements set limits to both the environment and the number of animals, while depth imaging restricts the environment. While others have attempted to automate the annotation of mouse grooming using a machine learning classifier, available techniques are not robust for multiple animal coat colors, lighting conditions, and locations of the setup (van den Boom et al., 2017). Recent advances in computer vision also provide general purpose solutions for marker-less tracking in lab animals (Mathis et al., 2018; Pereira et al., 2019), which will provide richer features to extend these traditional machine learning techniques. However, the adoption of these techniques to mouse behavior has yet to be seen. Even so, human action detection leaderboards suggest that while the approach of pose estimation is powerful, it routinely underperforms end-to-end solutions that utilize raw video input for action classification (Feichtenhofer et al., 2019; Choutas et al., 2018).

Here, we use neural networks to directly classify mouse grooming behavior from images. Grooming represents a form of stereotyped or patterned behavior of considerable biological importance consisting of a range of small to large actions. Grooming is an innate behavior conserved across animal species, including mammals (Spruijt et al., 1992; Kaljeff et al., 2010). In rodents, a significant amount of waking behavior, between 20%-50%, consists of grooming (Van de Weerd et al., 2001; Spruijt et al., 1992; Bolles, 1960). Grooming serves many adaptive functions such as coat and body care, stress reduction, dearrowal, social functions, thermoregulation, nociception, as well as other functions (Spruijt et al., 1992; Kaljeff et al., 2010; Fentress, 1988). The neural circuitry that regulates grooming behavior has been studied, although much remains unknown. Importantly, grooming and other patterned behaviors are endophenotypes for many psychiatric illnesses. For instance, a high level of stereotyped behavior is seen in autism spectrum disorder (ASD), where as Parkinson’s disease shows an inability to generate patterned behaviors (Kaljeff et al., 2010). Therefore, the accurate and automated analysis of grooming behavior represents important progress in behavioral neuroscience. We also reason that successful development of a neural network architecture for grooming behavior classification will be transferable to other
behaviors by changing the training data.

We apply a general machine learning solution to mouse grooming and develop a classifier that performs at human level. This classifier performs across all 62 inbred and F1 hybrid strains of mice consisting of visually diverse coat colors, body shapes, and sizes. We explore reasons why our network has an upper limit on performance that seems to be concordant with human annotations.

Human level performance comes at a cost of a large amount of labeled training data. We identify environmental and genetic regulators of grooming behavior in the open field. Finally, we apply our grooming behavior solution to a genetically diverse mouse population and characterize the grooming pattern of the mouse in an open field. We use these data to carry out a genome wide association study (GWAS) and identify the genetic architecture that regulates heritable variation in grooming and open field behaviors in the laboratory mouse. Combined we propose a generalizable solution to complex action detection and apply it towards grooming behavior.

Results

Mouse Grooming

Behavior varies widely on both time and space scales, from fine spatial movements such as whisking, blinking, or tremors to large spatial movements such as turning or walking, and temporally from milliseconds to minutes. We sought to develop a classifier that could generalize to complex behaviors seen in the mouse. Grooming consists of syntaxes that are small or micro-motions (paw lick) to mid-size movements (unilateral and bilateral face wash) and large movements (flank licking).

Figure 1A. There are also rare syntaxes such as genital and tail grooming. The length of time of grooming can vary from sub-seconds to minutes.

Annotating Grooming

Our approach to annotating grooming classifies each frame in a video as the mouse being in one of two states: grooming or not grooming. We specify that a frame should be annotated as grooming when the mouse is performing any of the syntaxes of grooming, whether or not the mouse is performing a stereotyped syntactic chain of grooming. This includes a wide variety of postures and action duration which contribute to a diverse visual appearance. This also explicitly includes individual paw licks as grooming, despite isolated paw licks not constituting a bout of grooming. Scratching is excluded from being classified as grooming.

We investigated the variability in manual grooming annotation by humans by tasking 5 trained annotators with labeling the same six 5-minute videos (30 minutes total, Figure 1). To help human scorers, we provided these videos from a top-down and side view of the mouse (Figure 1B). These videos included C57BL/6j, BTBR and CAST/EiJ mouse strains. We gave each annotator the same instructions to label the behavior (see Methods). We observe strong agreement (89.1% average) between annotators, which is in concordance with prior work annotating mouse grooming behavior (Kyzar et al., 2011). To examine disagreements between annotators, we classified them into three classes: missed bout, skipped break, and misalignment (Figure 1–Figure Supplement 1). Missed bout calls are made when a disagreement occurs in a not-grooming call. Similarly, skipped break calls are made when a disagreement occurs in a grooming call. Finally, misalignment is called when both annotators agree that grooming is either starting or ending but disagree on the exact frame in which this occurs. The most frequent type of error is misalignment, accounting for 50% of total duration of disagreement frames annotated and 75% of the disagreement calls (Figure 1–Figure Supplement 1).

We next constructed a large annotation data set to train a machine learning algorithm. While most machine learning contests seeking to solve tasks similar to ours have widely varied data set sizes, we leveraged network performance in these contests for design of our data set. Networks in these contests perform well when an individual class contains at least 10,000 annotated frames (Girdhar et al., 2019). As the number of annotations in a class exceeds 100,000, network perfor-
**Figure 1.** Annotating mouse grooming behavior. (A) Mouse grooming contains a wide variety of postures. Paw licking, face-washing, flank linking, as well as other syntaxes all contribute to this visually diverse behavior. (B) Grooming ethograms for 6 videos by 5 different trained annotators. Overall, there is very high agreement between human annotators. (C) Quantification of the agreement overlap between individual annotators. Average agreement between all annotators is 89.13%.

**Figure 1-Figure supplement 1.** Additional details of annotator disagreements.
mance for this task achieves mean average precision (mAP) scores above 0.7 (Girdhar et al., 2019; Zhang et al., 2019). With deep learning approaches, model performance benefits from additional annotations (Sun et al., 2017). To ensure success, we set out to annotate over 2 million frames with either grooming or not grooming. We aimed to balance this data set for grooming behavior by selecting video segments based on tracking heuristics, prioritizing segments with low velocity because a mouse cannot be grooming while walking. We also cropped the video frame to be centered on the mouse to reduce visual clutter using our tracker (Geuthier et al., 2019). This cropping centered around the mouse follows the video tube approach, as seen in the current state of the art (Feichtenhofer et al., 2019). From a pool of 7 validated annotators, we obtained 2 annotations for 1,253 video segments totaling 2,637,363 frames with 94.3% agreement between annotations (Figure 2A).

**Proposed Neural Network Solution**

We trained a neural network classifier using our large annotated data set. Of the 1,253 video segments, we held out 153 video clips for validation. Using this split, we achieve similar distributions of frame-level classifications between training and validation sets (Figure 2A). Our machine learning approach takes video input data and produces an ethogram output for grooming behavior (Figure 2B). Functionally, our neural network model takes an input of 16112x112 frames, applies multiple layers of 3D convolutions, 3D pooling, and fully connected layers to produce a prediction for only the last frame (Figure 2C). To predict a completed ethogram for a video, we slide the 16-frame window across the video.

We compared our neural network approach to a previously established machine learning approach for annotating lab animal behavior, JAABA (Kabra et al., 2013). Our neural network achieves 93.7% accuracy and 91.9% true positive rate (TPR) with a 5% false positive rate (FPR) (Figure 3A). In comparison, the JAABA trained classifier achieves a lower performance of 84.9% accuracy and 64.2% TPR at a 5% FPR (Figure 3A). Due to memory limitations of JAABA, we could only train it using 20% of our training set. To test whether the training set size accounts for this poorer performance by JAABA, training our neural network using the same 20% of our training set still led to outperformance of JAABA (Figure 3B). Using different sized training data sets, we observe improved validation performance with increasing data set size (Figure 2–Figure Supplement 1). Using an interactive training protocol recommended by the authors of JAABA, we observe decreased performance. This is likely due to the drastic size difference of the annotated data sets used in training (475,000 frames vs 17,000 frames).

Our neural network approach is as good as human annotators, given our previous observations in Figure 1B-C of 89% agreement. We inspected the receiver operating characteristic (ROC) curve performance on a per-video basis and found that performance was not uniform across all videos (Figure 2–Figure Supplement 2). The majority of the 153 validation videos are adequately annotated by both the neural network and JAABA. However, 2 videos perform poorly with both algorithms and 7 videos show drastic improvement using a neural network over the JAABA trained classifier. Manual visual inspection of the 2 videos where both algorithms perform poorly suggests that they did not provide sufficient visual information to annotate grooming.

While developing our final neural network solution, we applied two forms of consensus modalities to improve single-model performance (Figure 3C). Each trained model makes slightly different predictions, due to being randomly initialized. By training multiple models and merging the predictions, we achieve a slight improvement on validation performance. Additionally, we also modified the input image for different predictions. Rotating and reflecting the input image appear visually different for neural networks. We achieve 32 separate predictions for every frame by training 4 models and applying 8 rotation and reflection transformations on the input. We merge these individual predictions by averaging the probability predictions. This consensus modality improves the ROC area under the curve (AUC) from 0.975 to 0.978. We attempted other approaches for merging the 32 predictions, including selecting the max value or applying a vote (median prediction). Averaging the
Figure 2. Neural Network based action detection. (A) Distribution of our mouse grooming data set and performance of machine learning algorithms applied to this data set. A total of 2,637,363 frames were annotated across 1,253 video clips by two different human annotators to create this data set. The outer ring represents the training data set distribution while the inner ring represents the validation data set distribution. (B) A visual description of the classification approach that we implement. To analyze an entire video, we pass a sliding window of frames into a neural network. (C) Our network takes video input and produces a grooming prediction for a single frame.
prediction probabilities achieved the best performance (Figure 3–Figure Supplement 3A). Finally, we apply a temporal smoothing filter over 46 frames of prediction. We identify 46 frames to be the optimal window for a rolling average (Figure 3–Figure Supplement 3B), which results in a final accuracy of 93.7% (ROC AUC of 0.984).

Our network can only make predictions on half a second of worth of information. To ensure our validation performance is indicative of the wide diversity of mouse strains, we investigated the extremes of grooming bout predictions in our large strain survey data set which was not annotated by humans. While most of the long bout (>2 minutes) predictions were real, there were some false positives in which the mouse was resting in a grooming-like posture. To mitigate these rare false positives, we implemented a heuristic to adjust predictions. We experimentally identified that grooming motion typically causes ellipse-fit shape changes (W/L) to have a standard deviation greater than $2.5 \times 10^{-4}$. When a mouse is resting, the shape changes (W/L) standard deviation does not exceed $2 \times 10^{-5}$. Knowing that a mouse’s posture in resting may be visually similar to a grooming posture, we assigned predictions in time segments where the standard deviation of shape change (W/L) over a 31 frame window is less than $5 \times 10^{-3}$ to a “not grooming” prediction. Of all the frames in this difficult to annotate posture, 12% were classified as grooming. This suggests that this is not a failure case for our network, but rather a limitation of the network when only given half a second worth of information to make a prediction.

This approach is capable of handling varying mouse postures as well as physical appearance, e.g. coat color and body weight. We observe good performance over a wide variety of postures and coat colors (Figure 3C-D, Supplementary Videos 1-9). Even nude mice, which have a drastically different appearance than other mice, achieve good performance. Visually, we observe instances where a small number of frame orientations and models make incorrect predictions. Despite this, the consensus classifier makes the correct prediction.

**Definition of Grooming Behavioral Metrics**

We designed a variety of grooming behavioral metrics that describe both grooming quantity and grooming pattern. Following prior work (Kalueff et al., 2010), we define a single grooming bout as continuous time spent grooming without interruption that exceeds 3 seconds. We allow brief pauses (less than 10s), but do not allow any locomotor activity for this merging of time segments spent grooming. Specifically, a pause occurs when motion of the mouse does not exceed twice its average body length. From this, we obtain a grooming ethogram for each mouse (Figure 4A). Using the ethogram, we sum the total duration of grooming calls in all grooming bouts to calculate the total duration of grooming. Once we have the number of bouts and total duration, we calculate the average bout duration by dividing the two. For measurement purposes, we calculate the 5-minute, 20-minute, and 55-minute summaries of these measurements. We included 5 and 20 minutes because these are typical open field assay durations.

Using 1-minute binned data, we calculate a variety of grooming pattern metrics (Figure 4B). We fit a linear slope to discover temporal patterning of grooming during the 55 minute assay (GrTimeSlope55min). Positive slopes for total grooming duration infer that the individual mouse is increasing its time spent grooming the longer it remains in the open field test. Negative slopes for total grooming duration infer that the mouse spends more time grooming at the start of the open field test than at the end. This is typically due to the mouse choosing to spend more time doing another activity over grooming, such as sleeping. Positive slopes for number of bouts infer that the mouse is initiating more grooming bouts the longer it remains in the open field test. Using 5-minute binned data, we designed additional metrics to describe grooming pattern by selecting which 5-minute bin a mouse spent the most time grooming (GrPeakMidBin) and the time duration spent grooming (GrPeakVal) in that minute. We also calculate a ratio between these values (GrPeakSlope).

Finally, when we look at strain-level averages of grooming, we identify how long a strain remains at its peak grooming (GrPeakLength).

We compared a variety of open field measurements including both grooming behavior and
**Figure 3.** Validation of neural network model. (A) Agreement between the annotators while creating the data set compared to the accuracy of the algorithms predicting on this data set. We compare the machine learning models against only annotations where the annotators agree. (B) Receiver operating characteristic (ROC) curve for 3 machine learning techniques trained on the training set and applied to the validation set. Our final neural network model approach achieves the highest area under curve (AUC) value of 0.9843402. (C) A visual description of our proposed consensus solution. We use a 32x consensus approach where we trained 4 separate models and gave 8 frame viewpoints to each. To combine these predictions, we average all 32 predictions. While one viewpoint from one model can be wrong, the mean prediction using this consensus improves accuracy. (D-E) Example frames where the model is correctly predicting grooming and not-grooming behavior. Also see Supplementary Videos 1-9.

**Figure 3-Figure supplement 1.** Additional ROC curve subsets.

**Figure 3-Figure supplement 2.** Validation performance of algorithm split by video.

**Figure 3-Figure supplement 3.** Comparison of different consensus modalities and temporal smoothing.
Figure 4. Grooming and open field behavioral metrics. (A) Example grooming ethogram for a single animal. Time is on the x-axis and a blue colored bar signifies that the animal was performing grooming behavior during that time. We calculate summaries at 5, 20, and 55 minute ranges. (B) A visual description of how we define our grooming pattern phenotypes. (C) A table summarizing the 24 behavioral metrics we analyze. We group the phenotypes into 4 groups, including grooming quantity, grooming pattern, open field anxiety, and open field activity.
classical open field measurements (Figure 4C). We separated these 24 phenotypes into 4 groups. Grooming quantity describes how much an animal grooms, while grooming pattern metrics describe how an animal changes its grooming behavior over time. Including traditional measurements, open field anxiety measurements are phenotypes that have been validated to measure anxiety. Open field activity describes the general activity level of an animal.

**Sex and Environment Covariate Analysis of Grooming Behavior**

With this trained classifier, we sought to determine whether sex and environment affect grooming behavior expression in an open field, specifically grooming duration. We used data collected over 29 months for two strains, C57BL/6j and C57BL/6Nj to carry out this analysis. These two strains are substrains that were identical in the year 1951 and are two of the most widely used strains in mouse studies (Bryant et al., 2018). C57BL/6j is the mouse reference strain and C57BL/6Nj has been used by the International Mouse Phenotyping Consortium (IMPC) to generate a large amount of phenotypic data (Brown and Moore, 2012). We analyzed 775 C57BL/6j (317F, 458M) and 563 C57BL/6Nj (240F, 323M) mice tested over two and a half years under a wide variety of experimental conditions and ages. Across all these novel exposures in an open field, we quantified their grooming behavior for the first 30 minutes (Figure 5, 669 hours total data). We analyzed the data for effect of sex, season, time of day, age, room origin of the mice, light levels, tester, and white noise. To achieve this, we applied a stepwise linear model selection to model these covariates. Both forward and backward model selection results match. After identifying significant covariates, we applied a second round of model selection that includes sex interaction terms. The model selection identified sex, strain, room of origin, time of day, and season as significant. In contrast, age, weight, presence of white noise, and tester are not significant under our testing conditions. Additionally, the interaction between sex and both room of origin and season were identified as significant covariates.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>&lt; 2.2 × 10^{-16} ***</td>
</tr>
<tr>
<td>Strain</td>
<td>0.0267546 *</td>
</tr>
<tr>
<td>RoomOrigin</td>
<td>5.357 × 10^{-13} ***</td>
</tr>
<tr>
<td>Morning</td>
<td>0.0001506 ***</td>
</tr>
<tr>
<td>Season</td>
<td>0.0039826 **</td>
</tr>
<tr>
<td>Sex by RoomOrigin</td>
<td>0.0001568 ***</td>
</tr>
<tr>
<td>Sex by Season</td>
<td>0.0235954 *</td>
</tr>
</tbody>
</table>

We found an effect of strain (Figure 5A, \( p = 0.0268 \) C57BL/6j vs C57BL/6Nj) on grooming duration. Although the effect size is small, C57BL/6Nj groom more than C57BL/6j. Additionally, we observe a sex difference (Figure 5A, \( p < 2.2 × 10^{-16} \) males vs females). Males groomed more than females in both strains. Since sex has a strong effect, we included interaction terms with other covariates in a second pass of our model selection. The model identified season as a significant covariate (Figure 5B, \( p = 0.004 \)). Surprisingly, the model also identified an interaction between sex and season (\( p = 0.024 \)). Female mice for both strains show a increase in grooming during the summer and a decrease in the winter. Males do not show this trend, visually confirming the sex-season interaction.

We carried out testing between 8AM and 4PM. To determine if the time of test affects grooming behavior, we split the data into two groups: morning (8am to noon) and afternoon (noon to 4pm). We observe a clear effect of time of day (Figure 5C, \( p = 0.00015 \)). Mice tested in the morning groom more overall. We tested mice of different ages, ranging from 6 weeks to 26 weeks old. At the beginning of every test, we weighed the mice and found them to have a range of 16g to 42g. We did not observe any significant effect of age (Figure 5D, \( r = -0.065, p = 0.119 \)) or body weight (\( r = 0.206, p = 0.289 \)) on grooming duration, although we did not test “old” mice, generally considered to be more than 18 months old.

We compared the grooming levels of mice that were shipped from production rooms in a nearby building at our institution to our testing room with mice bred and raised in a room adjacent to the
testing room (B2B). Six production rooms supplied exclusively C57BL/6j (AX4, AX29, AX1, MP23, MP14, MP15), 3 rooms supplied exclusively C57BL/6NJ (MP13, MP16, AX5), and one room supplied both strains (AX8). All shipped mice were housed in B2B for at least a week prior to testing. We observe a significant effect for room of origin (Figure 5E, $p = 5.357 \times 10^{-11}$). For instance, C57BL/6j males from AX4 and AX29 were low groomers compared to other rooms, including B2B. Shipped C57BL/6NJ from all rooms seem to have low levels of grooming compared with B2B. We conclude that room of origin and shipping can both have effects on grooming behaviors.

We tested two light levels, 350-450 lux and 500-600 lux white light (5600K). We observe significant effects of light levels on grooming behavior (Figure 5F, $p = 0.04873$). Females from both strains groom more in lower light, however males don’t seem to be affected. Despite this, our model did not include a light-sex interaction, suggesting that other covariates better account for the visual interaction with sex here.

The open field assays were carried out by one of two male testers, although the majority of tests were carried out by tester 2. Both testers carefully followed a testing protocol designed to minimize tester variation, which only involves weighing the mouse and placing the mouse into the arena. We observed no significant effect (Figure 5G, $p = 0.65718$) between testers.

Finally, white noise is often added to open field assays in order to create a uniform background noise levels and to mask noise created by the experimenter (Gould, 2009). Although the effects of white noise have not been extensively studied in mice, existing data indicate that higher levels of white noise increase ambulation (Weyers et al., 1994). We tested the effects of white noise (70db) on grooming behavior of C57BL/6j and C57BL/6NJ mice and found no significant difference in duration spent grooming. Although there appears to be a stratification present for both C57BL/6j and C57BL/6NJ females, other cofactors better account for this.

Combined, these results indicate that environmental factors such as season, time of day, and room origin of the mice affect grooming behavior and may serve as environmental confounds in any grooming study. We also investigated age, body weight, light level, tester, and white noise and found these cofactors to not influence grooming behavior under our experimental conditions.

**Strain Differences for Grooming Behavior**

Next, we used the grooming classifier to carry out a survey of grooming behavior in the inbred mouse. We tested 43 classical laboratory and 8 wild derived strains and 11 F1 hybrid mice from the Jackson Laboratory (JAX) mouse production facility. These were tested over a 31 month period and in most cases consisted of a single mouse shipment of mice from JAX production. Other than C57BL/6j and C57BL/6NJ, on average we tested 8 males and 8 females of an average age of 11 weeks for each strain. Each mouse was tested for 55 minutes in the open field as previously described (Geuther et al., 2019). This data set consisted of 2457 animals and 2252 hours of video. Video data were classified for grooming behavior as well as open field activity and anxiety metrics. Behavior metrics were extracted as described in Figure 4. In order to visualize the variance in phenotypes, we plotted each animal across all strains with corresponding strain mean and 1 standard deviation range and ethograms of select strains Figure 6. We distinguish between classical laboratory strains and wild derived inbred strains.

**Grooming amount and pattern in genetically diverse mice** We observed large continuous variance in total grooming time, average length of grooming bouts, and the number of grooming bouts in the 55-minute open field assay (Figure 6). Total grooming time varied from 2-3 minutes in strains such as 129X1/SvJ and BALB/cByJ to 12 minutes in strains such as SJL/J and PWD/PhJ. Strains such as 129X1/SvJ and C57BR/cdj have less than 10 bouts, whereas MA/MyJ had almost 40 bouts. The bout duration also varied from 5 seconds to approximately 50 seconds in BALB/cByJ and PWD/PhJ, respectively. In order to visualize relationships between phenotypes we created strain mean and 1SD range correlation plots (Figure 7). There is a positive correlation between the total grooming time and the number of bouts as well as the total grooming time and average bout duration. Overall, strains with high total grooming time have increased number of bouts as well as...
Figure 5. Sex and environmental covariate analysis of grooming behavior (total time grooming) in the open field for C57BL/6j and C57BL/6Nj strains. Effect of sex (A), season (B), time of day (C), age (D), room of origin (E), light level (F), tester (G), and white noise (H).
longer duration of bouts. However, there does not seem to be a relationship between number of bouts and the average bout duration, implying that the bout lengths stay constant regardless of how many may occur (Figure 7). In general, C57BL/6j and C57BL6/NJ fall roughly in the middle for classical inbred strains.

We investigated the pattern of grooming over time by constructing a rate of change in 5 minute bins for each strain (Figure 8). There appeared to be visual structures in the data, so we used k-means to identify clusters. We identified three clusters of grooming patterns based on total grooming level and relative changes in 5-minute binned grooming data (Figure Supplement 1). Type 1 consists of 13 strains with an "inverted U" grooming pattern. These strains escalate grooming quickly once in the open field, reach a peak, and then start to decrease the amount of grooming, usually leading to a negative overall grooming slope. Often, we find animals from these strains are sleeping by the end of the 55-minute open field assay. These strains include both high groomers such as CZECHII/Eij, MOLF/Eij, and low groomers such as 129X1/SvJ and I/LnJ. Type 2 consists of 12 strains that are high grooming strains and do not reduce grooming by the end of the assay. They reach peak grooming early (e.g. PWD/PhJ, SJL/J and BTBR) or late (e.g. DBA/2J, CBA/J) and then remain at or near this peak level for the remainder of the assay. The defining feature of this group is that a high level of grooming is maintained throughout the assay. Type 3 consists of most of the strains (30) and shows steady increase in grooming until the end of the assay. Overall, the strains in this group are medium to low groomers with a constant low positive or flat slope. We conclude that under our experimental conditions there are at least three broad, albeit continuous, classes of observable grooming patterns in the mouse.

Wild derived vs. classical strain grooming patterns We compared grooming patterns between classical and wild derived laboratory strains. Classical laboratory strains are derived from limited genetic stock originating from Japanese and European mouse fanciers (Keeler, 1931; Morse, 1978; Silver, 1995). Classical laboratory inbred mouse lines represent the genome of Mus musculus domesticus (M.m. domesticus) 95% and Mus musculus musculus 5% (Yang et al., 2011). New wild derived inbred strains were established specifically to overcome the limited genetic diversity of the classical inbred lines (Guénet and Bonhomme, 2003; Koide et al., 2011). We observed that most wild derived strains groom for significantly higher duration and have longer average bout length than the classical inbred strains. Five of the highest 16 grooming strains are wild derived (PWD/PhJ, WSB/Eij, CZECHII/Eij, MSM/MsJ, MOLF/Eij in Figure 6A). The wild derived strains also have significantly longer bouts of grooming, with 6 of 16 longest average grooming bout strains from this group. Both the total grooming time and average bout length are significantly different between classical and wild-derived strains (Figure Supplement 1). These high grooming strains represent M.m. domesticus and M.m. musculus subspecies, which are the precursors to classical laboratory strains (Yang et al., 2011). These wild derived strains also represent much more of the natural genetic diversity of the mouse populations than the larger number of classical strains we tested. This leads us to conclude that the high levels of grooming seen in the wild derived strains better represent the normal levels of grooming behavior in mice. This implies that low grooming behavior may have been selected for in classical laboratory strains, at least as observed in our experimental conditions.

BTBR grooming pattern We also closely examined the grooming patterns of the BTBR strain, which has been proposed as a model with certain features of autism spectrum disorder (ASD). ASD is a complex neurodevelopmental disorder leading to repetitive behaviors and deficits in communication and social interaction (Association et al., 2013). Compared to C57BL/6j mice, BTBR have been shown to have high levels of repetitive behavior, low sociability, unusual vocalization, and behavioral inflexibility (McFarlane et al., 2008; Silverman et al., 2010; Moy et al., 2007; Scattoni et al., 2008). Repetitive behavior is often assessed by self grooming behavior, and drugs with efficacy in alleviating symptoms of repetitive behavior in ASD also reduce grooming in BTBR mice without affecting overall activity levels, which provides some level of construct validity (Silverman et al., 2012; Amodeo et al., 2017).
Figure 6. Strain survey of grooming phenotypes with representative ethograms. (A) Strain survey results for total grooming time. Strains present a smooth gradient of time spent grooming, with wild derived strains (purple) showing enrichment on the high end. (B) Representative ethograms showing strains with high and low total grooming time. (C) Strain survey results for number of grooming bouts. (D) Comparative ethograms for two strains with different number of bouts, but similar total time spent grooming. (E) Strain survey results for average grooming bout duration. (F) Comparative ethograms for two strains with different average bout length, but similar total time spent grooming.

Figure 6–Figure supplement 1. Grooming in wild-derived vs. classical inbred lines.
Figure 7. Relatedness of grooming phenotypes. Points indicate strain-level means. Lines indicate 1SD. (A) Strain survey comparing total grooming time and number of bouts. Wild derived strains and BTBR show enrichment for having high grooming but low bout numbers. (B) Strain survey comparing total grooming time and average bout duration. Strains that groom more also tend to have a longer average bout length. (C) Strain survey comparing number of bouts to average bout duration.
**Figure 8.** Three classes of grooming patterns in the open field revealed by clustering. Grooming duration in 5-minute bins is shown over the course of the open field experiment (blue line) and data from individual mice (grey points).

**Figure 8–Figure supplement 1.** K-means clustering of grooming patterns
We found that total grooming time in BTBR is high compared with C57BL/6j but is not exceptionally high compared to all strains (Figure 6), or even among classical inbred strains. C57BL/6j mice groom approximately 55 minutes over a 55 minute open field session, whereas BTBR groom approximately 12 minutes (Figure 6A). Several classical inbred strains have similar levels of high grooming, including SJ/LJ, DBA/1J, and CBA/Caj. The grooming pattern of BTBR belongs to Type 2 which contains 11 other strains (Figure 8). One distinguishing factor of BTBR mice is that they have longer average bouts of grooming from an early point in the open field (Figure 6 Figure Supplement 1). However, again they are not exceptionally high in average bout length measure (Figure 6 Figure Supplement 1). Strains such as SJ/LJ, PWD/PHj, MOLF/EiJ, NZB/BINj have similar long bouts from an early point. We conclude that BTBR display high levels of grooming with long grooming bouts, however, this behavior is similar to several wild derived and classical laboratory inbred strains and is not exceptional. Since we did not measure social interaction and other salient features of ASD, we do not argue against BTBR as an ASD model. In addition to BTBR, perhaps other strains from the type 2 group could serve as models of ASD.

Grooming mouse GWAS

Next we wished to understand the underlying genetic architecture of complex mouse grooming behavior and open field behaviors, and to relate these to human traits. We used the data from the 51 inbred strains and 11 F1 hybrid strains to carry out a genome wide association study (GWAS).

We did not include the 8 wild derived strains because they are highly divergent and can skew mouse GWAS analysis. We analyzed the 24 phenotypes categorized into four categories - open field activity, anxiety, grooming pattern, and quantity (Figure 4). We used a linear mixed model (LMM) implemented in Genome-wide Efficient Mixed Model Association (GEMMA) for this analysis in order to control for spurious association due to population structure (Zhou and Stephens, 2012). We first calculated heritability of each phenotype by determining the proportion of variance in phenotypes explained by the typed genotypes (PVE) Figure 9A. Heritability ranged from 6% to 68%, with 22/24 traits depicting heritability estimates greater than 20%, a reasonable estimate for behavioral traits in mice and humans (Valdar et al., 2006; Bouchard Jr, 2004), making them amenable for GWAS analysis Figure 9A.

We analyzed each phenotype using GEMMA, considering the resulting Wald test p-value. In order to correct for the multiple SNPs we are testing (222,966), and to account for the correlations between SNPs genotypes, we obtained an empirical threshold for the p-values by shuffling the values of one normally distributed phenotype (OFDistTraveled20m) and taking the minimal p-value of each permutation. This process resulted in a p-value threshold of $1.4 \times 10^{-5}$ that reflects a corrected p-value of 0.05 (Belmonte and Yurgelun-Todd, 2001). We defined quantitative trait loci (QTL) in the following manner: adjacent SNPs that have correlated genotypes ($r^2 > = 0.2$) were clustered together in a greedy way, starting with the SNP with the lowest p-value in the genome, assigning it a locus, adding all correlating SNPs and then moving forward to the next SNP with the lowest p-value until all the significant SNPs are assigned to QTL. The genetic architecture of inbred mouse strains dictates large linkage disequilibrium (LD) blocks (Figure 9B), resulting in QTL that span millions of base-pairs and contain multiple genes <Supp table>.

GWAS analysis of each phenotype resulted in 2 to 22 QTL (Figure Supplement 1). Overall, the open field activity had 15 QTL, anxiety 10, grooming pattern 76 and grooming quantity 51 QTL, leading to 130 QTL combined over all the tested phenotypes (Figure 9C). We observed pleiotropy with the same loci significantly associated with multiple phenotypes. Pleiotropy is expected since many of our phenotypes are correlated and individual traits may be regulated by similar genetic loci. For instance, we expected pleiotropy for grooming time in 55 and 20 minutes (GrTime55 and GrTime20) since these are correlated traits. We also expected that some loci regulating open field activity phenotypes may regulate grooming. In order to better understand the pleiotropic structure of our GWAS results, we generated a heat map of significant QTL across all phenotypes. We then clustered these, to find sets of QTL that regulate groups of phenotypes (Figure 9D). The phenotypes
were clustered into 5 subgroups consisting of grooming pattern (I), open field activity (II), open field anxiety (III), grooming length (IV), and grooming number and amount (V) (Figure 9D top x-axis). We found seven clusters of QTL that regulate combinations of these phenotypes (Figure 9D y-axis). For instance, clusters A and G are composed of pleiotropic QTL that regulate grooming length (IV) and grooming time but QTL in cluster G also regulate bout number and amount (V). QTL cluster D regulates open field activity and anxiety phenotypes. Cluster E contains QTL that regulate grooming and open field activity and anxiety phenotypes, but most of the SNPs only have significant p-values for either open field phenotypes or grooming phenotypes but not both, indicating that independent genetic loci are largely responsible for these phenotypes. We colored the associated SNPs in the Manhattan plot (Figure 9C) with colors to mark one of these seven QTL clusters (Figure 9D). These clusters ranged from 13 to 35 QTL, with the smallest being cluster F which is mostly pleiotropic for grooming number, and the largest cluster, cluster G, is pleiotropic for most of the grooming related phenotypes.

These highly pleiotropic genes included several genes known to regulate mammalian grooming, striatal function, neuronal development, and even language. Mammalian Phenotype Ontology Enrichment showed “nervous system development” as the most significant module with 178 genes (p = 7.5 × 10^{-4}) followed by preweaning lethality (p = 3.5 × 10^{-3}, 189 genes) and abnormal embryo development (p = 5.5 × 10^{-3}, 62 genes) <Supp table>. We carried out pathway analysis using KEGG and Reactome databases (Ogris et al., 2016). This analysis showed 14 disease pathways that are enriched including Parkinson's (9.68 × 10^{-4}), Huntington's 1.07 × 10^{-6}, non-alcoholic fatty liver disease (9.31 × 10^{-4}), and Alzheimer's (1.15 × 10^{-6}) as the most significantly enriched. Enriched pathways included oxidative phosphorylation (6.42 × 10^{-6}), ribosome (0.00000102), RNA transport (0.00000315), and ribosome biogenesis (0.00000465). Reactome enriched pathways included mitochondrial translation termination and elongation (2.50 × 10^{-19} and 5.89 × 10^{-19}, respectively), and ubiquitin-specific processing proteases (1.86 × 10^{-4}). The most pleiotropic gene was Sox5 which associated with 11 grooming and open field phenotypes. Sox5 has been extensively linked to neuronal differentiation, patterning, and stem cell maintenance (Lefebvre, 2010). Its dysregulation in humans has been implicated in Lamb-Shaffer syndrome and ASD, both neurodevelopmental disorder (Kwan, 2013; Zawerton et al., 2020). 102 genes were associated with 10 phenotypes, and 105 genes were associated with 9 phenotypes. We limited our analysis to genes with at least 6 significantly associated phenotypes, resulting in 860 genes. Other genes include FoxP1, which has been linked to striatal function and regulation of language (Bowers and Konopka, 2012). Ctnmb1, a regulator of Wnt signaling, and Grin2b, a regulator of glutamate signaling. Combined, this analysis indicates genes known to regulate nervous system function and development, and genes known to regulate neurodegenerative diseases as regulators of grooming and open field behaviors. The GWAS also begins to define the genetic architecture of grooming and open field behaviors in mice.

**PheWAS**

Finally, we wanted to link genes that are associated with open field and grooming phenotypes in the mouse with human phenotypes. We hypothesize that common underlying genetic and neuronal architectures exist between mouse and human, however, they can give rise to disparate phenotypes in each organism. For example, the misregulation of a certain pathway in the mouse may lead to over-grooming phenotypes but in humans the same pathways perturbation may manifest itself as neuroticism or obsessive compulsive disorder. These relationships between phenotypes and organisms can be revealed through identification of common underlying genetic architectures. In order to carry out such cross-species association and link the mouse genetic circuit of grooming to human phenotypes, we conducted a phenome-wide association study (PheWAS) with the Psychiatric Genetics GWAS catalog. First, we identified human orthologs of the 860 mouse grooming and open field genes with at least 6 degrees of pleiotropy. For each human ortholog, we downloaded PheWAS summary statistics from gwasATLAS (https://atlas.ctglab.nl/). Release 2: v20190117 (Watanabe et al., 2019). The gwasATLAS currently contains 4756 GWAS from 473 unique studies across 3302
unique traits which are classified into 28 domains. We only focused on the association in the Psychiatric domain with gene-level p value \( \leq 0.001 \). Second, in order to visualize and cluster these associations, we represented the relationships between genes and psychiatric traits by a weighted bipartite network, in which the width of an edge between a gene node and a Psychiatric trait node is proportional to the association strength (-log10(p value)). The size of a node is proportional to the number of associated genes or traits and the color of a trait node corresponds to the subchapter level in the Psychiatric domain. To identify modules within this network, we applied an improved community detection algorithm for maximizing weighted modularity in weighted bipartite networks (Dormann and Strauss, 2014). This analysis resulted in 8 gene-phenotype modules (Figure 10).

These modules contained between 15 and 32 individual phenotypes and between 41 and 103 genes. At the subchapter level, modules are enriched for temperament and personality phenotypes, mental and behavioral disorders (schizophrenia, bipolar, dementia), addiction (alcohol, tobacco, cannabinoid) obsessive-compulsive disorder, anxiety, and sleep.

Surprisingly, we found orthologs that show high levels of pleiotropy in mouse GWAS and the resulting human PheWAS. FOXP1 is the most pleiotropic gene with 35 human phenotypic associations, while SOX5 is second with 33 associations. In order to prioritize candidate modules for further research, first, we produced a ranked list of modules by calculating modularity score of each module, where high-ranking modules represent most promising candidates for further research (Newman and Girvan, 2004). Modularity scores of the modules ranged from 0.028 (module 8) to 0.103 (module 1). Second, we used Simes’ test to combine the p values of genes to obtain an overall p value for the association with each Psychiatric trait. Then the median of association (-log10(Simes p value)) was calculated in each detected module for prioritization. Using this method, module 1 again ranked at the top of eight modules (median = 5.29) (Figure 10–Figure Supplement 1).

Module 1 is primarily composed of temperament and personality phenotypes, including neuroticism, mood swings, and irritability traits. Genes in this module have a high level of pleiotropy in both human PheWAS and mouse GWAS. Eight of the 10 most pleiotropic genes from the PheWAS analysis are in this module. Genes in this module include SOX5 noted above, RANGAP1 with 31 associations, and EP300 with 23 significant human phenotypic associations. In conclusion, PheWAS analysis links conserved genes that regulate mouse grooming and other open field behaviors to human phenotypes. These human phenotypes include personality traits, addiction, and schizophrenia. We also find that the same genes are highly pleiotropic in mouse GWAS and human PheWAS analysis.

Discussion

Grooming is an ethologically conserved, neurobiologically important behavior that is often used as an endophenotype for psychiatric illnesses. It is a prototypical stereotyped, patterned behavior with highly variable posture and temporal length. Tools to automatically quantify behaviors such as grooming are needed by the behavioral research community and can be leveraged to gain insight into other complex behaviors (Spruijt et al., 1992; Kalueff et al., 2010, 2016). We present a neural network approach towards automated vertebrate model organism behavioral classification and ethogram generation that achieves human-level performance. This approach is simple, scalable, and can be carried out using standard open field apparatus, and is expected to be of use to the behavioral neuroscience community. Using this grooming behavior classifier, we analyze a large data set consisting over 2200 hours of video from dozens of mouse strains. We show environmental effects on grooming, patterns of grooming behavior in the laboratory mouse, carry out a mouse GWAS, and a human PheWAS to understand the underlying human-relevant genetic architecture of grooming and open field behavior in the laboratory mouse.

While the machine learning community has implemented a wide variety of solutions for human action detection, few applications have been applied to animal behavior. This may be in part due to the wide availability of human action data sets and the stringent performance requirements for human bio-behavioral research. We observe that the cost of achieving this stringent performance is
**Figure 9.** GWAS analysis of grooming and open field behaviors. (A) Heritability (PVE) estimates of the computed phenotypes. (B) LD blocks size - average genotype correlations for SNPs in different genomic distances. (C) GWAS results shown as a Manhattan plot of all of the phenotypes combined, colors are according to peak SNPs clusters (from D), all the SNPs in the same LD block are colored according to the peak SNP. Minimal p-value over all of the phenotypes for each SNP (D) Heatmap of all the significant peak SNPs for each. The each row (SNP) is colored according to the assigned cluster in the k-means clustering. The color from k-means cluster is used in C.

**Figure 9—Figure supplement 1.** Manhattan plot for individual phenotypes.
**Figure 10.** Human-Mouse trait relations through weighted bipartite network of PheWAS results. The width of an edge between a gene node (grey color) and a Psychiatric trait node is proportional to the association strength (-log10p value). The size of a node is proportional to the number of associated genes or traits and the color of a trait node corresponds to the subchapter level in the Psychiatric domain. Eight modules were visualized using Gephi 0.9.2 software.

**Figure 10–Figure supplement 1.** Box plot of Simes P values in eight modules.
very high, requiring a large quantity of annotations. More often than not, experimental paradigms are limited by cost to be short or small enough to cost less to allow manual annotation of the data. Machine learning approaches have been previously applied to automated annotation of behavioral data. We observe that our 3D convolutional neural network outperforms a JAABA classifier when trained on the same training data set. Our neural network achieves 91.9% true positive rate with a 5% false positive rate while the JAABA classifier achieves 64.2% true positive rate at a 5% false positive rate. This improvement makes the neural network solution usable for application to biological problems. This improvement is not uniform over all samples but is instead localized to certain types of grooming bouts. This suggests that although the JAABA classifier is powerful, it may be most useful for smaller and more uniform data sets. Experimental paradigms and behaviors with diverse expression will require a more powerful machine learning approach.

With the grooming classifier, we determined genetic and environmental factors that regulate this behavior. In a large data set collected over an 18-month period using two reference strains, C57BL/6j and C67BL/6N, we assessed effects on grooming of several fixed and dynamic factors including, sex, strain, age, time of day, season, tester, room origin, white noise, and body weight. All mice were housed in identical conditions for at least a week prior to testing. As expected, we observe strong effects of sex, time of day, and season. In many but not all studies, and not in the present study, tester effects have been observed in the open field in both mice and rats (Walsh and Cummins, 1976; McCall et al., 1969; Bohlen et al., 2014; Lewejohann et al., 2006). A recent study demonstrated that male experimenters or even clothes of males elicit stress responses from mice leading to increased thigmotaxis (Sorge et al., 2014).

To our surprise, we find that room of origin had a strong effect on grooming behavior of C57BL/6j and C57BL/6N. We did not see a clear directionality of effect between shipping mice and those bred in an adjacent room, and in some cases the effect size was high (z > 1). We hypothesize that this may be due to room-specific stress which has previously been demonstrated to alter grooming (Kalueff et al., 2016). Presumably, all external mice had similar experience of shipping from the production rooms to the testing area where they were housed identically for at least one week prior to testing. Thus, the potential differential stress experience was in the room of origin where the mouse was born and held until shipping. It is important to note that this shipping was only across buildings on the same campus, and shipping that involves air freight may have more drastic effects. This is a point of caution for use of grooming behavior as an endophenotype. Although we acclimated shipped mice to at least one week in an adjacent holding room prior to testing, a longer acclimation period may be required prior to testing.

We carried out a large strain survey to characterize and account for genetic diversity in grooming behavior in the laboratory mouse. We find three types of grooming patterns under our test conditions. Type 1 consists of mice that escalate and deescalate grooming within the 55-minute open field test. Strains in this group are often sleeping by the end of the assay, indicating a low arousal state towards the end of the assay. We hypothesize that these strains use grooming as a form of successful de-arousal, a behavior that has been previously noted in rats, birds, and apes (Spruijt et al., 1992; Delius, 1970). Similar to Type 1, Type 2 groomers escalate grooming quickly to reach peak grooming, however, this group does not deescalate grooming during our assay. We hypothesize that these strains need a longer time or may have a deficiency in de-arousal under our test conditions. Type 3 strains escalate for the duration of the assay indicating they have not reached peak grooming under our assay conditions. BTBR is a member of the Type 2 group with prolonged high levels of grooming from an early point, perhaps indicating a hyperaroused state, or an inability to de-arouse. BTBR mice have previously been shown to have high arousal states and altered dopamine function which may lead to the sustained high levels of grooming (Squillace et al., 2014). We postulate that other strains in the Type 2 grooming class may also show phenotypic features of ASD, warranting further study of ASD-related phenotypes in these strains.

Wild derived strains have distinct patterns of grooming compared to classical strains. Wild-derived strains groom significantly more and have longer grooming bouts. In our grooming
clustering analysis most of the wild derived strains belong to Type 1 or 2, where as most classical
strains belong in Type 3. In addition to *M. m domesticus*, the wild derived inbred lines we tested
represent *M. m musculus, M. m castaneous, and M. m molossinus* subspecies. Even though there
are dozens of classical inbred strains, there are approximately 5 million SNPs between any two
classical inbred laboratory strains such as C57BL/6J and DBA2J (*Keane et al., 2011*). Indeed, over
97% of the genome of classical strains can be explained by fewer than ten haplotypes indicating
small number of classes within which all strains are identical by descent with respect to a common
ancestor (*Yang et al., 2011*). Classical laboratory strains are derived from mouse fanciers in China,
Japan and Europe before being co-opted for biomedical research (*Morse, 1978; Silver, 1995*). Wild
derived inbred strains such as CAST/Eij and PWK/Phj have over 17 million SNPs compared to B6j,
and WSB/Eij have 8 million SNPs. Thus, the 7 wild derived strains we tested represent far more of
the genetic diversity present in the natural mouse population than the numerous classical inbred
laboratory strains. Behaviors seen in the wild-derived strains is more likely to represent behaviors
in the natural mouse population.

Mouse fanciers breed mice for visual and behavioral distinctiveness, and many exhibit them in
competitive shows. Mouse fanciers judge mice on "condition and temperament" and suggest that "it
us useless to show a mouse rough in coat or in anything but the mouse perfect condition" (*Davies,
1912*). Much like dogs and horses, the "best individuals should be mated together regardless of
relationship as long as mice are large, hardy, and free from disease" (*Davies, 1912*). It is plausible
that normal levels of grooming behavior seen in wild mice was considered unhygienic or indicative
of parasites such as lice, tics, fleas, or mites. High grooming could be interpreted as poor condition
and would lead the mouse fancier to select mouse strains with low grooming behaviors. This
selection could account for low grooming seen in the classical strains.

We used the strain survey data to conduct a mouse GWAS which identified 130 QTL that
regulates heritable variation in open field and grooming behaviors. We find that the majority of the
grooming traits are moderately to highly heritable. A previous study using the BXD recombinant
inbred panel identified 1 significant locus on chromosome 4 that regulates grooming and open
field activity (*Delprato et al., 2017*). We closely analyzed 862 genes in the QTL interval that are
highly pleiotropic and find enriched pathways that regulate neuronal development and function.
We then associated regions belong to one of 7 clusters which regulate combinations of open
field and grooming phenotypes. Mouse grooming can be used as a model of human grooming
 disorders such as trichotillomania, however, grooming is regulated by the basal ganglia and other
brain regions and can be used more broadly as an endophenotype for many psychiatric traits,
including ASD, schizophrenia, and Parkinson's (*Kalueff et al., 2016*). We conducted a PheWAS with
the highly pleiotropic genes and identified human psychiatric traits that are associated with these
genes. This approach allows us to link mouse and human phenotypes through the underlying
genetic architecture. This approach links human temperament, personality traits, schizophrenia,
and bipolar disorder to mouse phenotypes. Future research is needed to definitively link mouse
genetic variants to altered behavior. Our GWAS results are a starting point for understanding the
genetic architecture of grooming behavior and will require functional studies in the future to assign
causation.

In conclusion, we describe a neural network based machine learning approach for action
detection in mice and apply it towards grooming behavior. Using this tool we characterize grooming
behavior and its underlying genetic architecture in the laboratory mice. Our approach to grooming
is simple and can be carried out using standard open field apparatus and should be of use to the
behavioral neuroscience community.
Materials and Methods

Animals

All animals were obtained from The Jackson Laboratory production colonies or bred in a room adjacent to the testing room as previously described (Geuther et al., 2019; Kumar et al., 2011). All behavioral tests were performed in accordance with approved protocols from The Jackson Laboratory Institutional Animal Care and Use Committee guidelines.

Data set annotation

We selected data to annotate by training a preliminary JAABA classifier for grooming, then clipping video chunks based on predictions for a wide variety of videos. The initial JAABA classifier was trained on 13 short clips that were manually enriched for grooming activity. This classifier is intentionally weak, designed simply to prioritize video clips that would be beneficial to select for annotation. We clipped video time segments with 150 frames surrounding grooming activity prediction to mitigate chances of a highly imbalanced data set. We generated 1,253 video clips which total 2,637,363 frames. Each video has variable duration, depending upon the grooming prediction length. The shortest video clip contains 500 frames, while the longest video clip contains 23,922 frames. The median video clip length is 1,348 frames.

From here, we trained 7 annotators. From this pool of 7 trained annotators, we assigned two annotators to annotate each video clip completely. If there was confusion for a specific frame or sequence of frames, we allowed the annotators to request additional opinions. Annotators were required to provide a “Grooming” or “Not Grooming” annotation for each frame, with the intent that difficult frames to annotate would get different annotations from each annotator. We only train and validate using frames in which annotators agree, which reduces the total frames to 2,487,883.

Neural network model

Our neural network follows a typical feature encoder structure except using 3D convolution and pooling layers instead of 2D. We start with a 16x112x112x1 input video segment, where 16 refers to the time dimension of the input and 1 refers to the color depth (monochrome). Each convolution layer that we apply is zero-padded to maintain the same height and width dimension. Additionally, each convolution layer is followed by batch normalization and ReLU activation. First, we apply two sequential 3D convolution layers with a kernel size of 3x3x3 and number of filters of 4. Second, we apply a max pooling layer of shape 2x2x2 to result in a new tensor shape of 8x64x64x4. We repeat this two 3D convolution and max pool, doubling the filter depth each time, an additional 3 more times which results in a 1x8x8x32 tensor shape. We apply two final 3D convolutions with a 1x3x3 kernel size and 64 filter depth, resulting in a 1x8x8x64 tensor shape. Here, we flatten the network to produce a 64x64 tensor. After flattening we apply two fully connected layers, each with 128 filter depth, batch normalization, and ReLU activations. Finally, we add one more fully connected layer with only 2 filter depth and a softmax activation. This final layer is used as the output probabilities for not grooming and grooming predictions.

Neural network training

We trained 4 individual neural networks using the same training set and 4 independent initializations. During training, we randomly sample video chunks from the data set where the final frame contains an annotation where the annotators agree. Since we sample a 16 frame duration, this refers to the 16th frame’s annotation. If a frame selected does not have 15 frames of video earlier, the tensor is padded with 0-initialized frames. We apply random rotations and reflections of the data, achieving an 8x increase in effective data set size. The loss function we use in our network is a categorical cross entropy loss, comparing the softmax prediction from the network to a one-hot vector with the correct classification. We use the Adam optimizer with an initial learning rate of $10^{-5}$. We apply a decay schedule of learning rate to halve the learning rate if 5 epochs persist without validation.
accuracy increase. We also employ a stop criteria if validation accuracy does not improve by 1% after
10 epochs. During training, we assemble a batch size of 128 example video clips. Typical training
would be done after 13-15 epochs, running for 23-25 epochs without additional improvement.

**JAABA Training**

We trained JAABA classifiers using two different approaches. Our first approach was using the
guidelines provided by the software developers. This involves interactively and iteratively training
classifiers. The data selection approach is to annotate some data, then prioritize new annotations
where the algorithm is unsure or incorrectly making predictions. We continued this interactive
training until the algorithm no longer made improvements in a k-fold cross validation.

Our second approach was to subset our large annotated data set to fit into JAABA and train
on the agreeing annotations. Initially, we attempted to utilize the entire training data set, but our
machine did not have enough RAM to handle the entire training data set. The workstation we used
contained 96 GB of available RAM. We wrote a custom script to convert our annotation format to
populate annotations in a JAABA classifier file. To confirm our data was input correctly, we examined
the annotations from within the JAABA interface. Once we created this file, we could simply train the
JAABA classification using JAABA's interface. After training, we applied the model to the validation
data set to compare with our neural network models. We repeated this with various sizes of training
data sets.

**Definition of Grooming Behavioral Metrics**

Here we describe a variety of grooming behavioral metrics that we use in following analyses.
Following the approach that (Kalueff et al., 2010) used, we define a single grooming bout as a duration of
continuous time spent grooming without interruption that exceeds 3 seconds. We allow brief pauses (less than 10 s), but do not allow any locomotor activity for this merging of
time segments spent grooming. Specifically, a pause occurs when motion of the mouse does not
exceed twice its average body length. In order to reduce the complexity of the data, we summarize
the grooming duration, number of bouts, and average bout duration into 1-minute segments. In
order to have a whole number of bouts per time duration, we assign grooming bouts to the time
segment when a bout begins. In rare instances where multiple-minute bouts occur, this allows for a
1-minute time segment to contain more than 1-minute worth of grooming duration.

From here, we sum the total duration of grooming calls in all grooming bouts to calculate the
total duration of grooming. Note that this excludes un-joined grooming segments less than 3 s
duration as they are not considered a bout. Additionally, we count the total number of bouts. Once
we have the number of bouts and total duration, we calculate the average bout duration by dividing
the two. Finally, we bin the data into one minute time segments and fit a linear line to the data.

Positive slopes for total grooming duration infer that the individual mouse is increasing its time
spent grooming; the longer it remains in the open field test. Negative slopes for total grooming
duration infer that the mouse spends more time grooming at the start of the open field test than at the
end. This is typically due to the mouse choosing to spend more time doing another activity over
grooming, such as sleeping. Positive slopes for number of bouts infer that the mouse is initiating
more grooming bouts the longer it remains in the open field test.

**Data collection and reporting**

Protocols for data collection were previously described in (Geuther et al., 2019). In brief, each
animal was video recorded from a top-down viewpoint for 55 minutes of novel open field exposure.
No power analysis was used sample size for C57BL/6J vs C57BL/6J data since this data is longitudi-
nal control data. Power analysis for the strain survey data showed that with 16 animals (8M/8F), we
have 80% power to detect a effect size of 1 (Cohen’s d). Outliers in the strain survey were removed
when individuals measured a value of $GrTime_{55m} < Q_1 - 1.5 \times IQR$ or $GrTime_{55m} > Q_3 + 1.5 \times IQR$,
where $Q_1$ is the first quartile, $Q_3$ is the third quartile, and $IQR$ is the interquartile range.
All behavioral data will be available in the Mouse Phenome Database (MPD), and code and models will be available in Kumar Lab Github account (https://github.com/KumarLabJason and kumarlab.org).

**Genome wide association study (GWAS)**

The phenotypes obtained by the machine learning algorithm for several strains were used to study the association between the genome and the strains behaviour. A subset of ten individuals from each combination of strain and sex were randomly selected from the tested mice to ensure equal within group sample sizes. The genotypes of the different strains were obtained from the mouse phenome database (https://phenome.jax.org/genotypes). The Mouse Diversity Array (MDA) genotypes were used, di-allele genomes were deduced from parent genomes. SNPs with at least 10% MAF and at most 5% missing data were used, resulting with 222,967 SNPs out of 470,818 SNPs genotyped in the MDA array. LMM method from the GEMMA software package (Zhou and Stephens, 2012) was used for GWAS of each phenotype with the Wald test for computing the p-values. A Leave One Chromosome Out (LOCO) approach was used, each chromosome was tested using a kinship matrix computed using the other chromosomes to avoid proximal contamination. Initial results showed a wide peak in chromosome 7 around the Tyr gene, a well known coat-color locus, across most phenotypes. To control for this phenomenon, the genotype at SNP rs32105080 was used as a covariate when running GEMMA. Sex was also used as a covariate. To evaluate SNP heritability, GEMMA was used without the LOCO approach. The kinship matrix was evaluated using all the SNPs in the genome and GEMMA LMM output of the proportion of variance in phenotypes explained – the PVE and the PVESE were used as chip heritability and its standard error.

LD decay - To estimate the LD decay pairs of SNPs that are at most 2.5 Mbp apart had their genotypes correlation computed using Pearson correlation. The pairs of SNPs were divided into bins according to their distance, each bin size being 5,000 bp. The average correlation $r^2$ coefficient of pairs of SNPs in each bin were averaged and plotted against the average SNPs distance and smoothed using loess function. We chose $r^2$ of 0.2 to be the threshold to assign SNPs to the same QTL. QTL were determined for each phenotype GWAS by sorting the SNPs according to their p-values, then, for each SNP, determining a locus centered at this SNP by adding other SNPs with high correlation ($r^2 > 0.2$) to the peak SNP. A locus was limited to 10 million bp from the initial peak SNP selected. These regions were used to find proximate but uncorrelated SNPs in the genome.

The peak SNPs were aggregated from all the phenotypes and the p-values from all the phenotypes GWAS results were used to cluster the peaks into clusters using the k-means algorithm implemented in R. After observing the results, seven clusters were chosen.

To combine the 24 phenotypes tested, we took the phenotypes from the same group and all of the phenotypes and for each SNP took the minimal p-value from the phenotypes in the group.

The GWAS execution was wrapped in an R package called mouseGWAS available on github: https://github.com/TheJacksonLaboratory/mousegwas, it also includes a singularity container definition file and a nextflow pipeline for regenerating the results.

**Acknowledgments**

We thank Drs. Kristin Branson and Manyank Kabra (Janelia Research Campus) for providing initial directions on the project and help with JAABA. We thank members of the Kumar Lab for helpful advice and Taneli Helenius for editing. We thank JAX Information Technology team members Edward Zaborowski, Shane Sanders, Rich Brey, David McKenzie, and Jason Macklin for infrastructure support. This work was funded by The Jackson Laboratory Directors Innovation Fund, National Institute of Health DA041668 (NIDA), DA048634 (NIDA), and Brain and Behavioral Foundation Young Investigator Award (V.K.). This work used the National Science Foundation (NSF) Extreme Science and Engineering Discovery Environment (XSEDE) XStream service at Stanford University through allocation TG-DBS170004 (to V.K.). All code and training data will be available at Kumarlab.org and Kumar Lab Github.
Competing Interests

The authors have no competing interest.

References


Davies C. Fancy Mice: Their Varieties and Management as Pets Or for Show, Including the Latest Scientific Information as to Breeding for Colour. LU Gili; 1912.


Gould T. Mood and Anxiety Related Phenotypes in Mice, Neumethods 42. DOI 10.1007/978-1-60761-303-9_1, Humana Press, a part of Springer Science+…; 2009.


Figure 1–Figure supplement 1. (A) We classify the types of annotation disagreement into three classes of errors: missed bouts, misalignment, and skipped breaks. (B-C) We quantify the types of errors in Figure 1. (B) The sum of frames that fall into each error category. 37.5% of frames are missed bouts, 50% are misalignment, and 12.4% are skipped breaks. The types of errors are not uniformly distributed across annotators as annotator 2 accounts for the most missed bout frame counts, annotator 4 accounts for the most misalignment frame counts, and annotator 1 accounts for the most skipped break frame counts. (C) Counting the number of multi-frame occurrences of errors, we observe a similar distribution. 19.2% of call are missed bouts, 74.6% are misalignment, and 6.3% are skipped breaks.
**Figure 3-Figure supplement 1.** Performance of the network using different training data set sizes. As the number of training data samples increases, the ROC curve performance increases up to a point.
Figure 3—Figure supplement 2. Clusters of validation video ROC performance by machine learning approaches. (A) The majority of validation videos have good performance. (B) Some validation videos suffer from slightly degraded performance. (C) Validation videos where the JAABA approach performs slightly worse than our neural network approach. (D) Two validation videos showed poor performance using both machine learning approaches. Upon inspection of these videos, the annotated frames for grooming are visually difficult to classify. (E) 7 validation videos showed good performance using the neural network but a clear drop in performance using JAABA. All videos that contain no positive grooming annotated frames do not have a ROC curve and were excluded from this figure.
**Figure 3—Figure supplement 3.** (A) ROC performance using different consensus modalities. All consensus modalities provide approximately the same result. (B) Temporal filter analysis.
Figure 6—Figure supplement 1. (A) Total grooming time of wild derived and classical strains are significantly different (* p < 0.05, Mann-Whitney Test). (B) Wild derived lines have significantly longer grooming bouts (**) p < 0.01, Mann-Whitney Test). In both graphs, BtBR strain is indicated in green.

Figure 8—Figure supplement 1. Results from the k-means clustering. The first two principal components from this clustering accounts for 81.7% of variance. 3 clusters were identified.
**Figure 9—Figure supplement 1.** LMM results for each SNP genotype using Wald test for each of the phenotypes
Figure 10–Figure supplement 1. Simes test was used to combine the p values of genes to obtain an overall p value for the association of each Psychiatric trait. Box plot showed the distribution of Simes P values in eight modules in Figure 10.
GrTimeSlope55m

Chromosome

-log(P-value)

0 2 4 6

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X

Cfap293 Wedd1 Synj1