

## Supporting Information

Steele *et al.* 10.1073/pnas.0610779104.

### SI Methods

**Behavioral Definitions.** The software cannot recognize any action less than six frames (the recording speed used was 30 frames per second), so any behavior that is less than 6/30 of a second in duration is not scored as a separate behavior. Behaviors are scored based on the posture of the mouse and the position of its body parts. Algorithms are proprietary. The behavioral definitions, some of which are merged in our data analysis, are listed below in alphabetical order:

*Awaken:* The end of a bout of rest (i.e., any activity that occurs after rest would be scored as an awaken event)

*Chew:* Mouse moves into a vertically huddled position with paws in front of face. When the mouse is eating food at the food bin then comes down to chew, the coming down and rearing back up are implied in the “chew” behavior and are not scored as “rearing.”

*Distance Traveled:* Distance traveled measures only lateral movement ( $x$  plane) in the cage (based on the center of mass of the mouse), and does not measure distance in the  $y$  or  $z$  planes. Thus it is an underestimate of total distance traveled.

*Drink: Start of behavior:* mouth is level with drinking spout and then the drinking action proceeds. *End of behavior:* mouth withdrawn from drinking spout. Sniffing before drinking is labeled as drinking as long as the drinking action follows.

*Eat:* Mouse’s snout is in the plane of the food bin, head and body movement is minimal. The speed of movement determines whether this is a “sniff” or “eat” behavior, with “sniff” as a faster movement than “eat.” Moreover, if a mouse sniffs before eating, then this is considered part of eating behavior.

*Groom:* Mouse uses both front paws to clean himself and rubs them over body and face in a circular movement. This behavior can be detected by its repetitive movements.

*Hang upside down: Start of behavior:* both forelimbs and hindquarters are at or above the midpoint between the cage floor and wire rack and limbs are in contact with wire rack. *End of behavior:* both paws up and hindquarters at or below midpoint of the cage.

*Hang vertical: Start of behavior:* both forelimbs up on wire rack (either the food bin or the top of the cage) AND both hindlimbs up or suspended. *End of behavior:* as soon as one paw touches the ground regardless of whether one or two paws are on the wire rack.

*Jump: Start of behavior:* nose points AND THEN body extends AND THEN at least one hind paw leaves the ground. Body extension and forward propulsion may happen simultaneously. Jumps may be initiated from ground or reared up position. *End of behavior:* one OR both hind paws on the ground.

*Pause:* Minimum of 3 s of “rest-like stillness” to be considered as “pause” but not as long

as “rest,” which is 30 s.

*Remain low:* Any prolonged inactivity that is neither “rest” nor “pause,” or any other detectable behavior.

*Rearing:* Encompasses eight behaviors measured separately by HCS. *Rear up - start of behavior:* mouse points up and then front paws leave the floor. *End of behavior:* mouse squatting on hind legs, back is curved and mouse is half way to the full stretch of rear up. *Rear up from partially reared - start of behavior:* mouse squatting on hind legs, back is curved and mouse is half way to the full stretch of rear up. *End of behavior:* mouse standing on hind legs only, behavior ends when nose reaches highest point. *Rear up to partially reared - start of behavior:* nose points up and then at least one paw leaves the floor. *End of behavior:* mouse squatting on hind legs, back is curved and mouse is half way to the full stretch of rear up. *Come down - start of behavior:* mouse is standing on hind legs, and nose begins to descend after highest point reached. *End of behavior:* at least one front paw is on the ground. *Come down to partially reared - start of behavior:* mouse is standing on hind legs, and nose begins to descend after highest point reached. *End of behavior:* mouse squatting on hind legs, back is curved and mouse is half way to the full stretch of rear up. *Come down from partially reared - start of behavior:* mouse squatting on hind legs, back is curved and mouse is half way to the full stretch of rear up. *End of behavior:* at least one front paw is on the ground.

*Rest:* Mouse is in a nonreared nonhanging posture and not moving except for breathing, which is a sustained movement. A minimum of 30 s is required to be considered as “rest.” Although 30 s appears a short threshold for “rest,” once mice remain inactive beyond 30 s they generally “rest” for much longer periods of time. The average rest bout lasts for 590 seconds (from  $n = 8$  control mice at 2 m.p.i.).

*Sniff:* Body stationary but nose bobbing. This is a secondary behavior, only to be charted when mouse is not involved in other behaviors.

*Stretch:* Head extends and/or hind legs past normal position. The mouse extends its body to maximum length possible either horizontally or vertically and the back usually arches into a C shape especially when in vertical position. *Start of behavior:* head begins the extension forward. *End of behavior:* mouse returns to normal resting state.

*Turn:* *Start of behavior:* Nose changes direction, then body follows, must be at least a quarter turn to be counted. When mouse walks and turns at the same time, turn takes priority.

*Twitch:* Any movement occurring during rest. Rest must occur immediately before and after the twitch.

*Unassigned:* A combination of “unknown behavior” and “no data.” For “unknown behavior,” a behavior is not recognized by HCS; typically these are very fast, erratic, or atypical movements. During periods of “no data” the HCS software does not recognize the mouse, and therefore does not score behaviors. This typically occurs during light/dark cycle transitions when the video cameras have to readjust to changing lighting conditions and the image becomes blurry and the software loses track of the mouse.

*Walk*: At least three legs move and propel the mouse forward.

**HCS Software.** HCS uses proprietary software algorithms to separate foreground objects (mice being recorded) from the background of the HC. It further identifies the animals' body parts (such as head, ears, mouth, fore- and hindlimbs, back, and tail) and assigns postures to the animals, rather than reducing animals to points in space. Initially, as purchased, the system assigned behaviors to the mice based on defined sequences of identified postures and recorded those. To adapt the system for use in our disease models, we refined or modified some behavioral definitions. For example, what we term "rearing" is a sum of eight separate behavioral classes defined by the HCS software, "rear up," "rear up from partially reared," "rear up to partially reared," "remain partially reared," "remain reared up," "come down," "come down to partially reared," "come down from partially reared." We also worked with CleverSys to improve the detection accuracy of some behaviors by increasing the signal-to-noise ratio. This often involved varying lighting and settings parameters (e.g., component size thresholds [the minimal and maximal pixel threshold values for a mouse]) to enable the software to better distinguish the mouse from the background. Initially the software could not perform analysis on videos longer than 2-3 h, and by trouble-shooting several software bugs, we were able to analyze 24-h videos. We improved the accuracy of the software from  $< 50\%$  to  $> 80\%$  on average across behaviors, with several behaviors approaching 99% accuracy.

To use HCS over the entire diurnal cycle, we established successful conditions for light- and dark-phase video capture. The night recording capability was virtually untested and unused until we began our studies. In our longer, more behaviorally representative, recordings, we found that mice burrowed into the bedding, making it impossible for the software to recognize the animal after only a few minutes of recording in the dark. Thus we tested various light conditions and amounts of bedding to achieve a situation in which the mouse still had enough bedding for basic husbandry, but could also be observed by the camera over a period of hours during the night.

We also helped to test and identify hardware components required for high-throughput analysis. Initially, the software was designed to operate in a "one cage-one camera" setup, limiting the throughput of the system to a level that was not useful for academic or industrial studies. We recognized the need to create a higher-throughput system, and successfully attempted a four-cage four-camera setup. To accomplish this, we tested various video components (e.g., quads equipped for digital video, video cards capable of capturing large videos from the quads) and data-collection parameters (e.g., the highest-resolution video that we could get without bogging down the processors) to achieve reliable video analysis. Finally, we contributed to software development by identifying bugs and making suggestions for improving the user-friendliness of the software. As just one example, the software was initially mischaracterizing several unrelated behaviors as "eat" even though mice were clearly engaging in other completely unrelated behaviors, such as walk or jump. It was only after our modifications, achieved with several reiterations of algorithms for scoring behavior that we identified important differences in eating behaviors between HD and control mice. Since we began beta testing HCS, we have used  $> 15$  generations of the software, each of which was adapted based on our advice.

**Multiparameter Behavioral Analysis.** Because the number of predictors (behaviors) exceeds the number of observations (mice), there was the danger of overfitting the data, i.e., of finding rules that were tuned to random variations in the particular mice we observed; such rules may be accurate when applied to these mice but may not generalize well to other mice. We used  $L_1$  regularization to control overfitting during logistic regression.  $L_1$  regularization

tends to force the coefficients of redundant or less-informative predictors to be exactly zero and so, in effect, automatically selects a subset of behaviors to use as predictors (17). This makes the rules easier to interpret. This algorithm entails the use of maximum-likelihood estimation to find the coefficients of the predictors, subject to the constraint that the sum of the absolute values of the coefficients (their  $L_1$  norm) be less than some adjustable parameter  $t$ , after centering each predictor to have mean 0 and variance 1. The parameter  $t$  was chosen automatically according to the Bayesian information criterion (BIC), by using the R package `glmPath`. Before computing logistic regressions with `glmPath`, raw behavior data (percent times for 17 behaviors, distance traveled, number of awaken events) were transformed via  $f(x) = \log(x + 0.01)$ . This, in most cases, made the behavior-specific variances for diseased and control mice more nearly equal and, in most cases, made the distribution of data for each class more nearly normal, based on maximum-likelihood estimates of the mean and variance. (Adding 0.01 avoids taking the log of zero.) Default parameters for `glmPath` were used, except for `max.arclength`, which was set to 0.05 so that more points along the regularization path would be computed exactly. Occasionally, at early time points the regularization parameter  $t$  with the best BIC score gave a rule with no nonzero coefficients, so that predictions fell exactly on the decision boundary. To break the ties in these cases during cross-validation, we used the next-largest value of  $t$  along the regularization path.

We assessed the robustness of the diagnoses using cross-validation, holding out two mice at a time (a diseased mouse and its littermate paired control), training a diagnostic rule on the remaining mice by using logistic regression with  $L_1$  regularization (17), and using this rule to predict (independently) whether each of the two removed mice were diseased or control. The number of correct predictions during cross-validation was significantly better than random for HD Tgs versus controls at week 6 and beyond, except for week 8, and for PrD mice vs. controls at 3.5 m.p.i. and subsequent time points ( $P < 0.05$ , exact binomial test, see SI Table 2). All predictions were correct for HD Tgs vs. controls beginning at 9 weeks, aside from a single misclassification at 12 weeks, and all predictions were correct for PrD mice vs. controls at 4.5 m.p.i. and beyond. Detailed results of cross-validation at all time points are given in SI Table 2. For HD mice at week 7, only one misclassification was made during cross-validation, of 14 mice total ( $P = 0.0009$ ), and for PrD mice at 4 m.p.i., only one misclassification was made during cross-validation, of 16 mice total ( $P = 0.0003$ ).

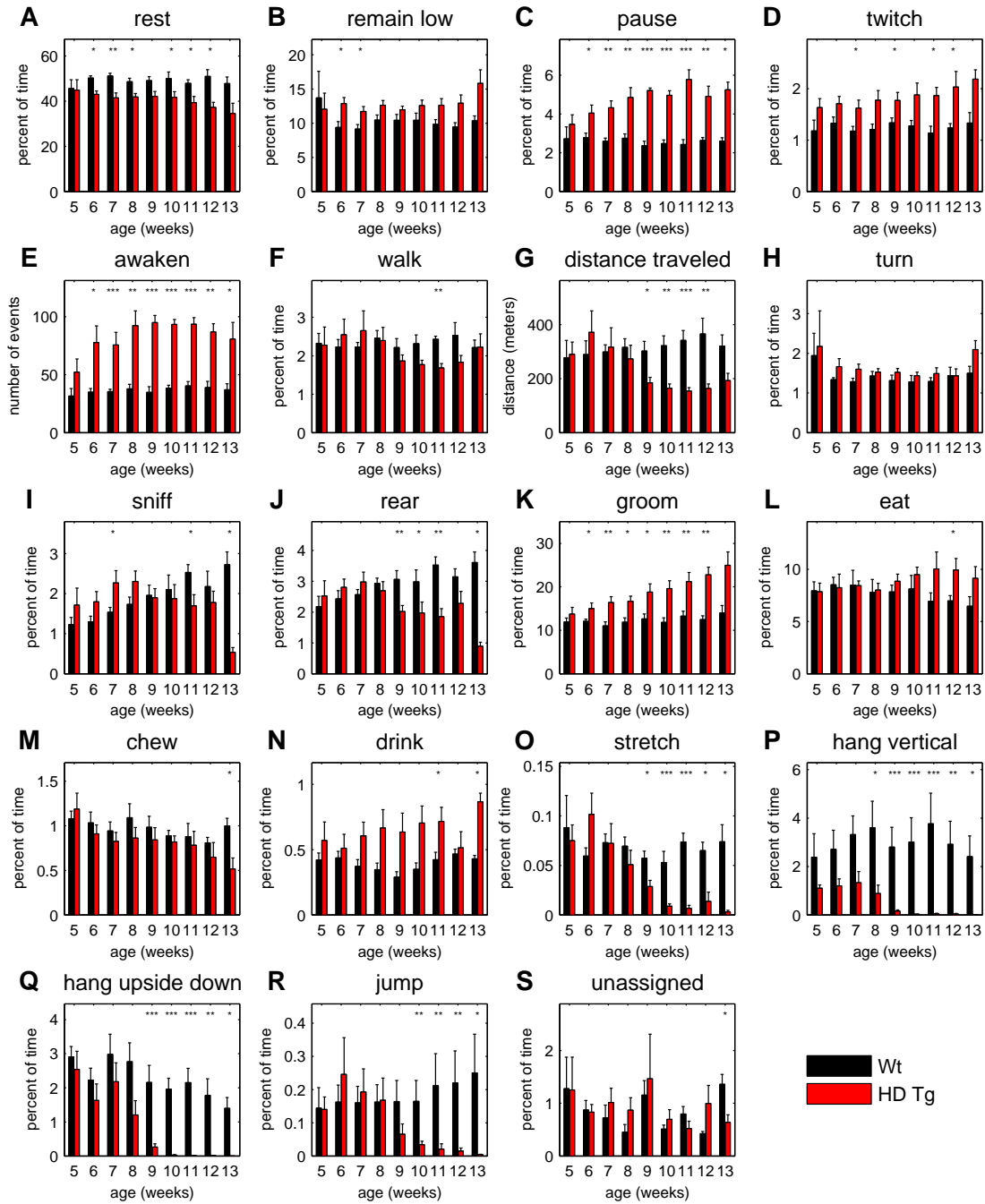


Figure 5: Behavioral alterations in Huntington's disease mice. Mean values ( $\pm$  SEM) are shown for rest (A), remain low (B), pause (C), twitch (D), awaken (E), walk (F), distance traveled (G), turn (H), sniff (I), rear (J), groom (K), eat (L), chew (M), drink (N), stretch (O), hang vertical (P), hang upside down (Q), jump (R), and unassigned behavior (S).  $P$  values were computed by using a two-tailed Wilcoxon rank-sum test (nonparametric).  $P$  values are indicated as follows \*,  $P < 0.05$ ; \*\*,  $P < 0.01$ ; and \*\*\*,  $P < 0.001$ . Sample sizes for each time point are as follows ( $n$  = number of mice) week 5  $n = 5$  HD Tg and WT control pairs, week 6  $n = 7$ , week 7  $n = 7$ , week 8  $n = 7$ , week 9  $n = 7$ , week 10  $n = 7$ , week 11  $n = 7$ , week 12  $n = 6$ , and week 13  $n = 4$ .

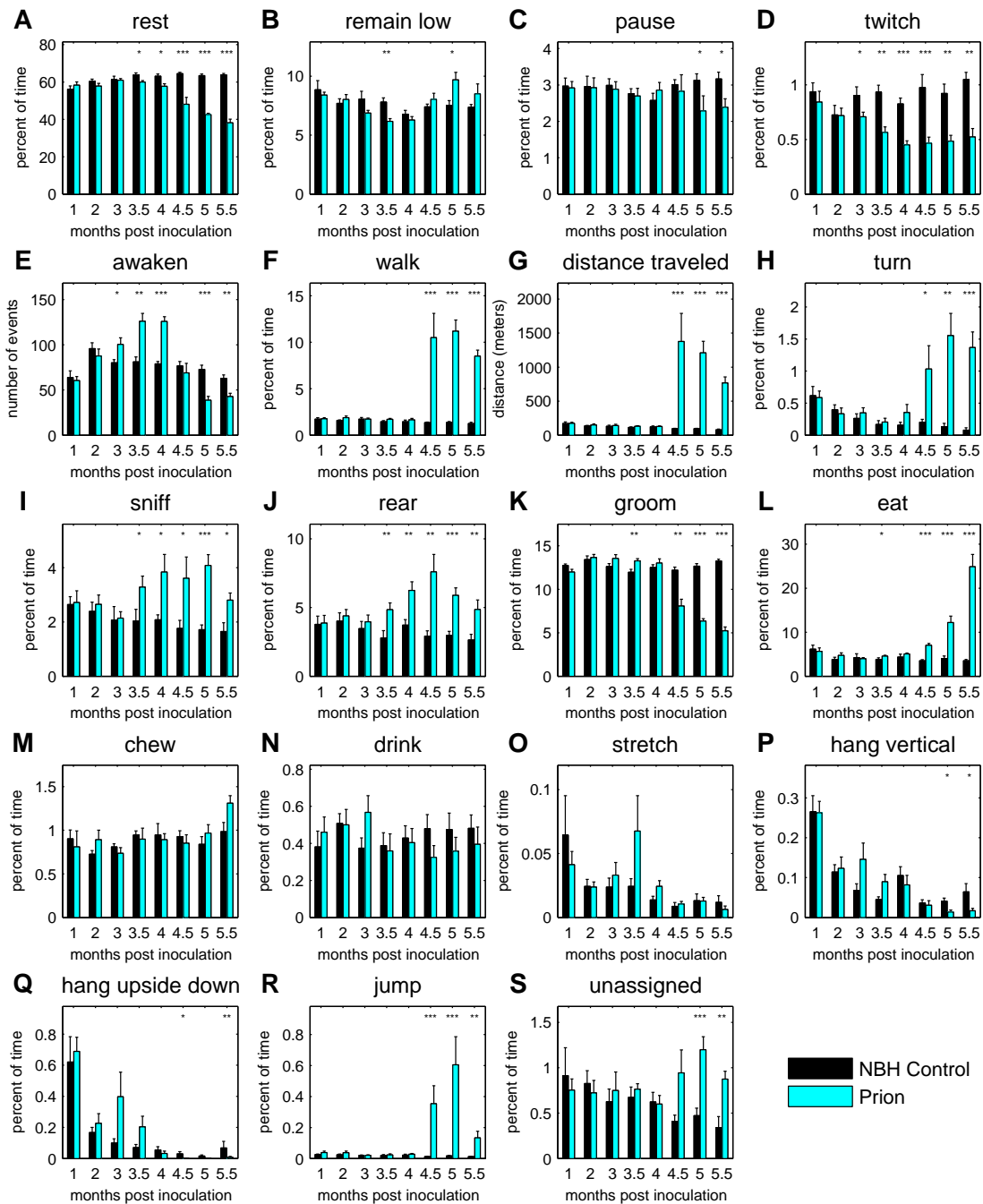


Figure 6: Behavioral alterations in prion disease mice. Mean values ( $\pm$  SEM) are shown for rest (A), remain low (B), pause (C), twitch (D), awoken (E), walk (F), distance traveled (G), turn (H), sniff (I), rear (J), groom (K), eat (L), chew (M), drink (N), stretch (O), hang vertical (P), hang upside down (Q), jump (R), and unassigned behavior (S). *P* values were computed by using a two-tailed Wilcoxon rank-sum test (nonparametric). *P* values are indicated as follows \*, *P* < 0.05; \*\*, *P* < 0.01; and \*\*\*, *P* < 0.001. Sample sizes for each time point are *n* = 8 prion and *n* = 8 normal brain homogenate inoculated controls at all time points except for 5 m.p.i. with *n* = 7 prion and *n* = 8 normal brain homogenate inoculated controls.

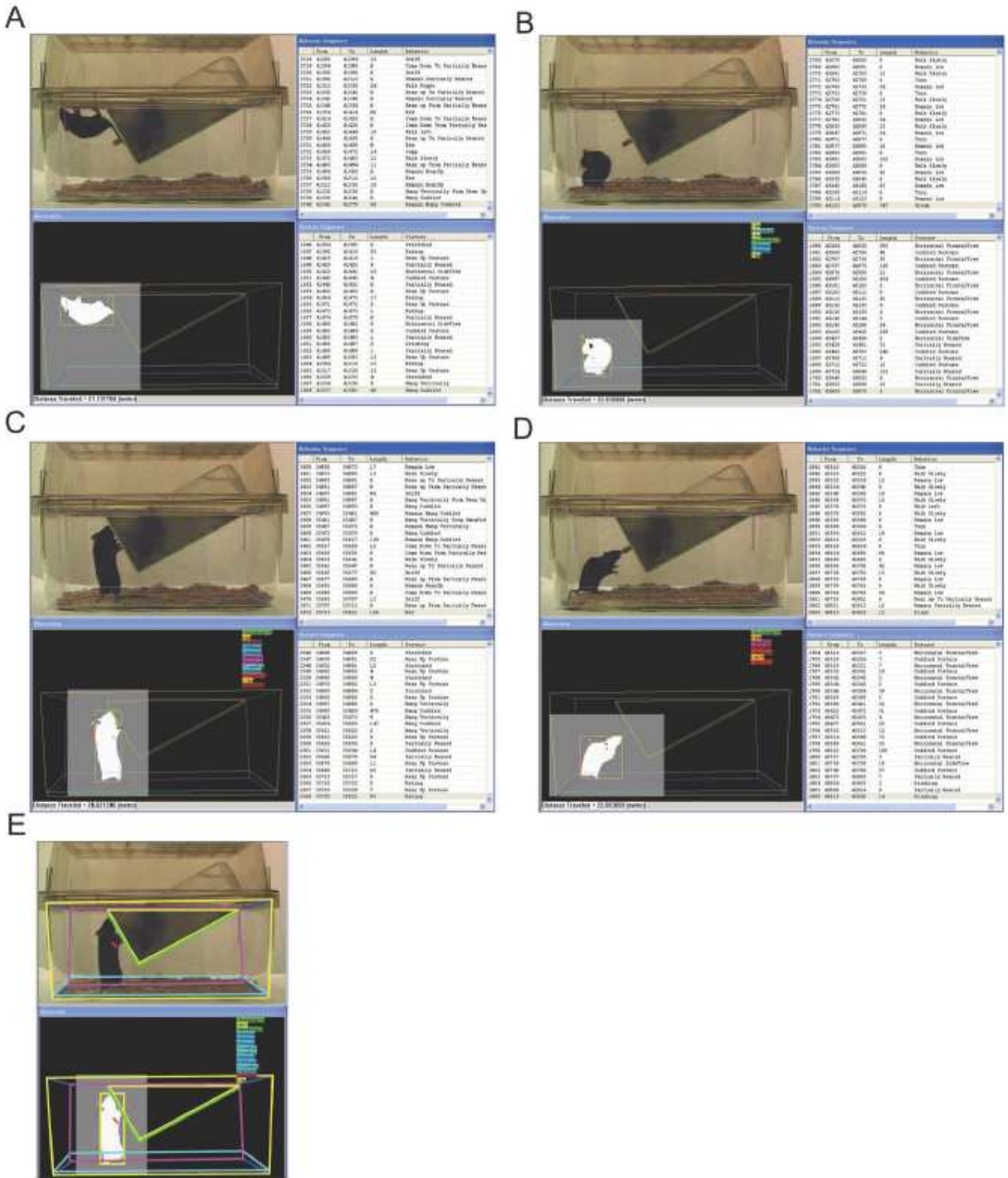


Figure 7: A picture of the video and HCS interface. Examples of hanging upside down (A), grooming (B), eating (C), and drinking (D) are shown. (E) The line weight of the image in C is increased to make the cage demarcations more clear.

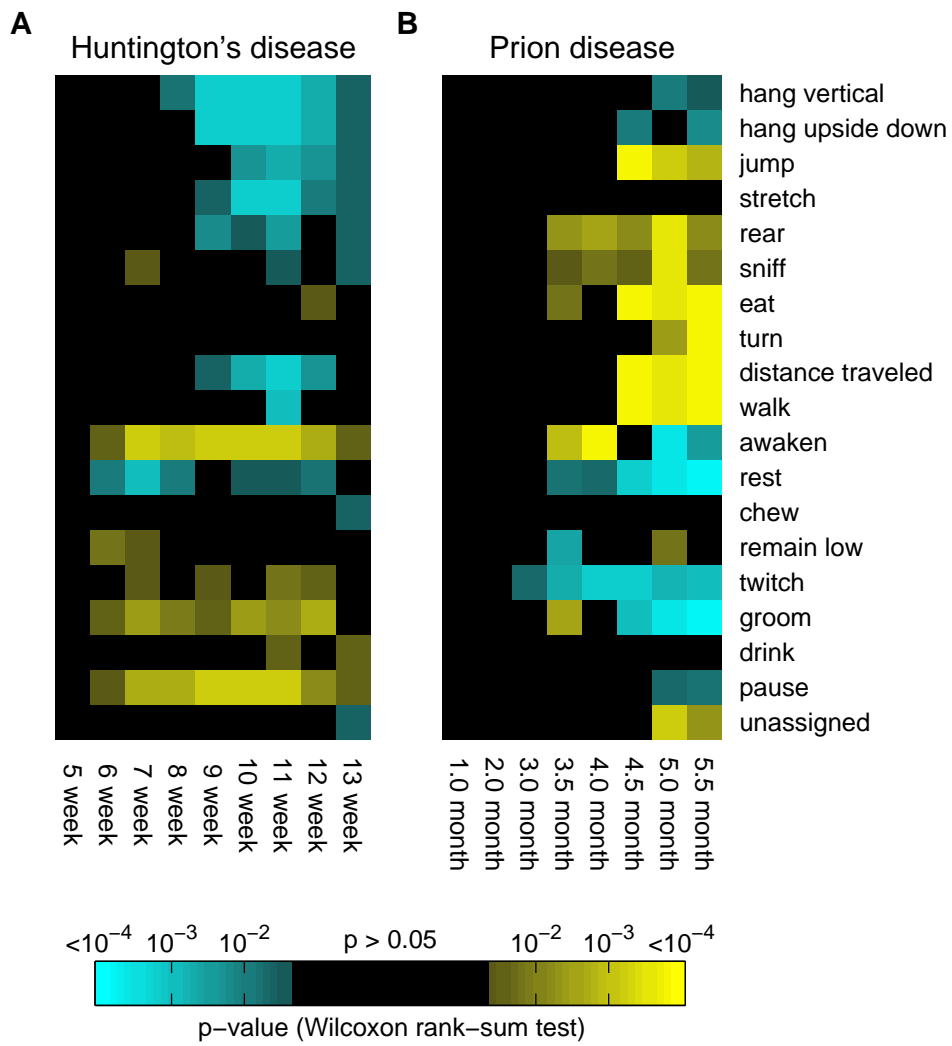


Figure 8: Fig. 3 represented in terms of  $P$  values (Wilcoxon rank-sum test).

Table 1: Accuracy assessment of scoring by HomeCageScan. The accuracy of HCS was assessed by inspecting  $\approx 100$  instances of each behavior (according to HCS) in a 24-h video of a single wild-type C57Bl/6 mouse. (For four behaviors,  $< 100$  total events were identified by HCS, so  $< 100$  were manually assessed.) The ratio of the number of the HCS instances that agreed with manual assessments (column 2) to the total number of scored instances (the sum of column 2 and column 3) gives the “specificity” of HCS for each behavior (column 6), i.e., the fraction of events that HCS classifies as a particular behavior that a human classifies as the same behavior. By weighting the incorrect classifications for each behavior (column 4) by the total number of HCS events for each behavior (column 5), we estimated a “confusion matrix” that would be obtained if we had manually assessed all  $\approx 30,000$  HCS events. From this we computed an estimate of the “sensitivity” of HCS for each behavior (column 7), i.e., the fraction of events that a human would classify as a particular behavior that HCS correctly identifies. (Note that this estimate is based on behavior start times as called by HCS rather than by a human.) We estimated the overall accuracy to be  $\approx 83\%$ , by taking the average of the specificities (column 6), weighted by the relative frequencies of the HCS events (column 5).

1	2	3	4	5	6	7
Behavior	#corr.	#incorr.	Incorrect should be:	#HCS events	Spec.	Sens.
Rest	67	3	remain low (2), groom (1)	70	0.96	1.00
Turn	91	9	groom (4), remain low (3), walk (1), unknown (1)	5185	0.91	0.94
Twitch	87	13	turn (7), sniff (3), groom (2), stretch (1)	785	0.87	0.99
Walk	95	5	turn (2), unknown (3)	4487	0.95	0.95
Jump	22	9	rear (9)	39	0.71	0.94
Rear	71	29	groom (15), eat (5), remain low (5), drink (2), chew (1), walk (1)	7259	0.71	0.92
Remain low	88	12	groom (7), turn (2), rear (2), walk (1)	7698	0.88	0.92
Eat	90	10	groom (6), rear (4)	1028	0.90	0.66
Drink	80	8	rear (8)	103	0.91	0.39
Chew	50	50	eat (26), rear (14), groom (7), hang vert. (3)	435	0.50	0.46
Groom	92	8	remain low (6), walk (1), chew (1)	580	0.92	0.19
Hang upside down	100	0	n/a	320	1.00	0.99
Hang vertical	75	25	rear (25)	570	0.75	0.94
Sniff	62	38	groom (18), chew (11), remain low (2)	1527	0.62	0.97
Sniff	62	38	rear (2), unknown (5)	1527	0.62	0.97
Awaken	62	6	twitch (6)	69	0.91	1.00
Stretch	90	10	hang vert. (6), eat (2), chew (1), hang upside down (1)	200	0.90	0.95
Pause	83	17	groom (12), chew (3), remain low (2)	219	0.83	1.00
Unknown (Unassigned)	37	63	groom (21), walk (17), turn (8), rear (8), eat (3), sniff (3), remain low (1), stretch (1), jump (1)	183	0.37	0.20

Table 2: Cross-validation of predictions. We assessed the robustness of the diagnoses using cross-validation, holding out two mice at a time (a diseased mouse and its littermate paired control), training a diagnostic rule on the remaining mice using logistic regression with  $L_1$  regularization (17), and using this rule to predict (independently) whether each of the two removed mice were diseased or control.  $P$  values were computed using the exact binomial test. Here, true-positive refers to a correct classification of a diseased mouse, true-negative to a correct classification of a control mouse, false-positive to an incorrect classification of a control mouse, and false-negative to a incorrect classification of a diseased mouse.

Huntington's Disease								
age (weeks)	true-pos	true-neg	false-pos	false-neg	correct	incorrect	total	binom p-val
5	2	3	2	3	5	5	10	0.62
6	5	6	1	2	11	3	14	0.029
7	6	7	0	1	13	1	14	0.00092
8	5	4	3	2	9	5	14	0.21
9	7	7	0	0	14	0	14	6.10E-05
10	7	7	0	0	14	0	14	6.10E-05
11	7	7	0	0	14	0	14	6.10E-05
12	5	6	0	1	11	1	12	0.0032
13	4	4	0	0	8	0	8	0.0039

Prion Disease								
age (m.p.i.)	true-pos	true-neg	false-pos	false-neg	correct	incorrect	total	binom p-val
1	5	5	3	3	10	6	16	0.23
2	2	4	4	6	6	10	16	0.89
3	4	4	4	4	8	8	16	0.60
3.5	6	6	2	2	12	4	16	0.038
4	7	8	0	1	15	1	16	0.00026
4.5	8	8	0	0	16	0	16	1.53E-05
5	7	7	0	0	14	0	14	6.10E-05
5.5	8	8	0	0	16	0	16	1.53E-05