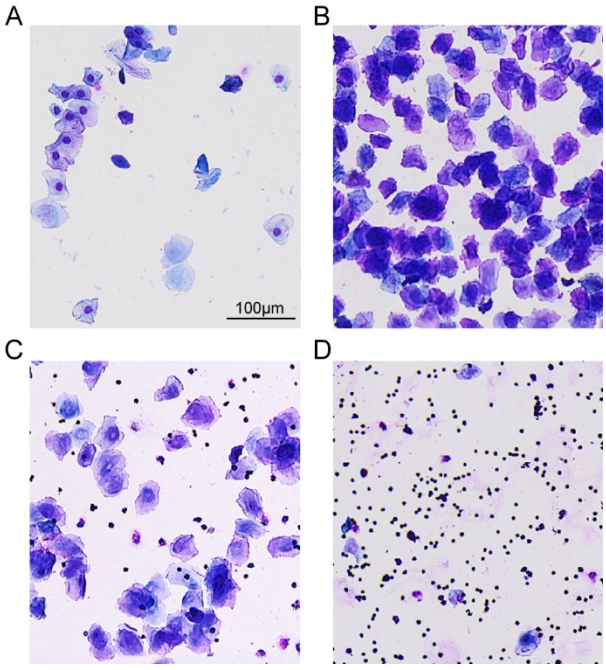
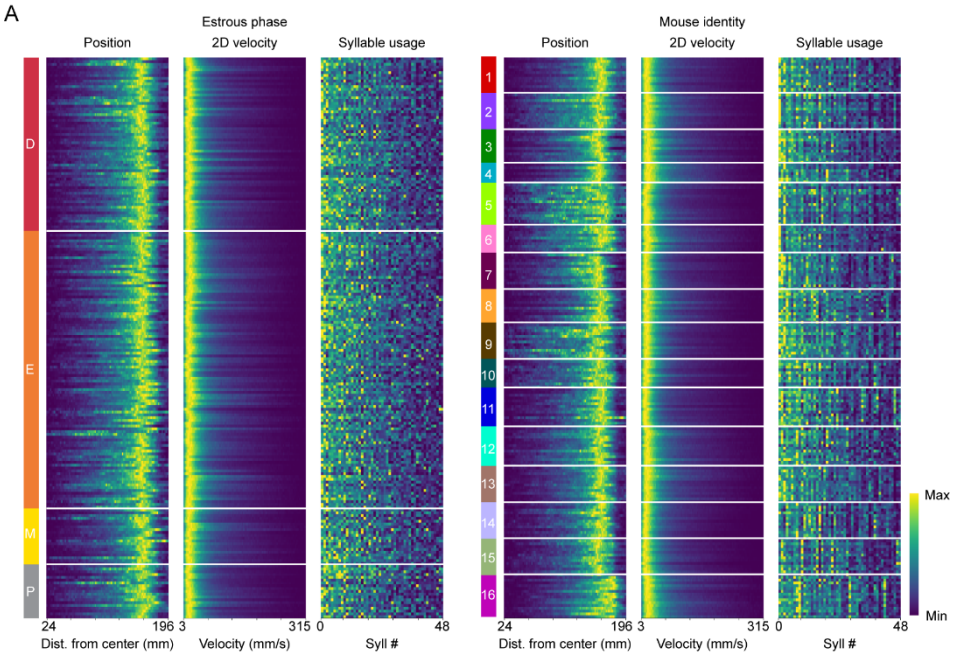
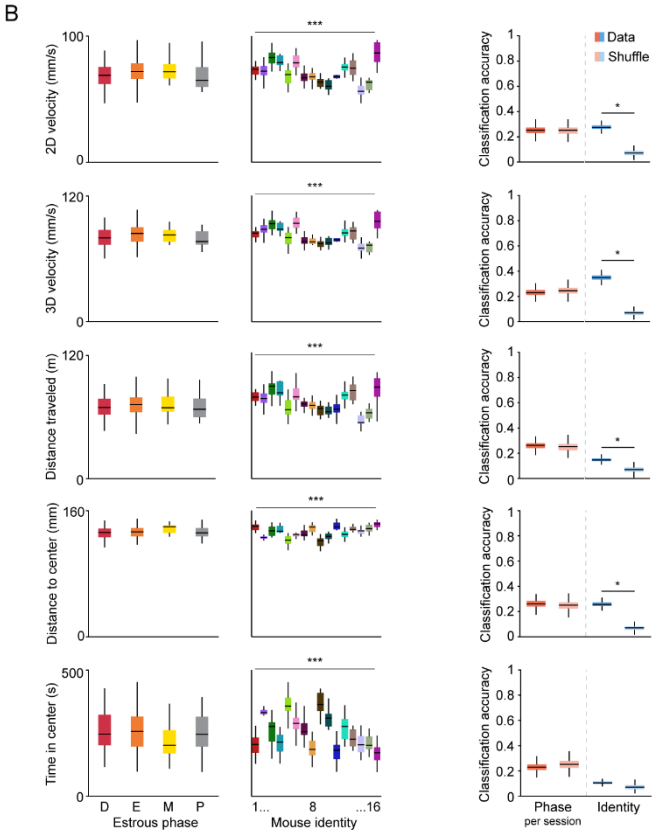
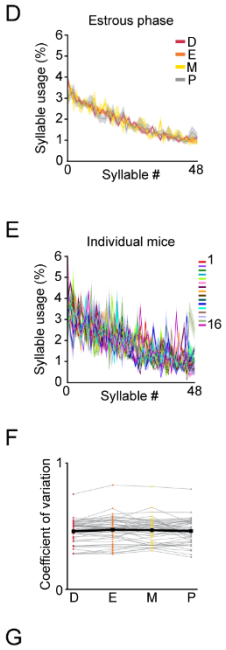


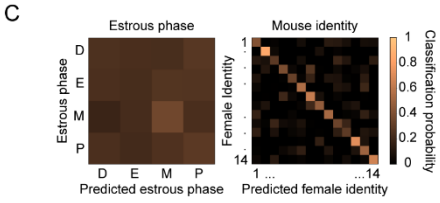
Figure S1: Identification of the four estrous phases, related to Figure 1.

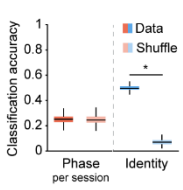
Representative images of vaginal smears used to define the four estrous stages in female mice: **A)** Proestrus; **B)** Estrus; **C)** Metestrus; **D)** Diestrus


A
Estrous phase
Position
2D velocity
Syllable usage
Mouse identity
Position
2D velocity
Syllable usage
D
E
M
P
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Max
Min
24
196
3
315
0
48
Dist. from center (mm)
Velocity (mm/s)
Syll #
Dist. from center (mm)
Velocity (mm/s)
Syll #


B
2D velocity (mm/s)
3D velocity (mm/s)
Distance traveled (m)
Distance to center (mm)
Time in center (s)
***
*
Data
Shuffle
Classification accuracy
D
E
M
P
Estrous phase
1...
8
...16
Mouse identity
Phase
Identity
per session


D
Estrous phase
Syllable usage (%)
Syllable #
E
Individual mice
F
Coefficient of variation
G
Mutual information (bits)
Phase
Identity
P:I


C
Estrous phase
Mouse identity
Estrous phase
Female Identity
Classification probability
Predicted estrous phase
Predicted female identity


Data
Shuffle
Classification accuracy
Phase
Identity
per session

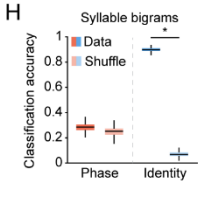

H
Syllable bigrams
Data
Shuffle
Classification accuracy
Phase
Identity

Figure S2: Estrous phase negligibly influences exploratory behavior, related to Figures 2.

A) Heatmaps depicting position, velocity, and syllable usages across all female dataset sorted by phase (left) or individual identity (right). For each measurement, values were binned into 49 bins (to match the number of syllables) and the colormap represents occupancy in each bin, normalized to max/min values for each session, and syllables are sorted by use across all sessions. White lines separate different phases or individuals. **B)** Left: Quantification of female behavior in the open field across estrous phases and in individual mice, as assessed via traditional kinematic scalar metrics. Kruskal-Wallis H-test was performed for all: for 2D velocity: $H_{phase(3)}$=5.56, p=0.13; $H_{individual(15)}$=128.803, p=3.64*10^{-20}. For 3D velocity: $H_{phase(3)}$=5.47, p=0.14; $H_{individual(15)}$=125.09, p=1.93*10^{-19}. For distance traveled: $H_{phase(3)}$=3.58, p=0.30; $H_{individual(15)}$=118.88, p=3.11*10^{-18}. For distance to center: $H_{phase(3)}$=5.27, p=0.15; $H_{individual(15)}$=128.44, p=4.28*10^{-20}. For time in center: $H_{phase(3)}$=4.19, p=0.24; $H_{individual(15)}$=133.66, p=4.06*10^{-21}. n sessions per phase: D=62, E=99, M=20, P=19; n sessions per individuals: 8-15. Right: Quantification of overall decoder performance for each scalar, matched to left panel. **C)** Left: Confusion matrix for classification accuracy of a decoder trained to predict estrous phase (left) or individual mouse identity (right) based on scalar quantiles of all scalars described in B concatenated. Right: Classification accuracy. **D-E)** Mean syllable usage distribution for phase **(D)** and individual mice **(E)**. Shaded area represents standard error of the mean. Kruskal–Wallis tests with Bonferroni correction were performed to assess differences in the use of each syllable between estrous phases (none significant). Syllables are sorted by use across all sessions, and the sorting is maintained for both panels. **F)** The overall variability of behavior over time during each session does not vary based upon estrous cycle. To assess this within-session variation, the coefficient of variation of syllable use during each 20 minute behavioral session was computed and graphed, with data binned into 5 minute chunks. Results are presented per syllable per phase, with gray lines representing individual syllables; Thick black line represents the mean across syllables per phase. Friedman test for repeated samples was performed. $X^2_{(3)}$=3.375, p=0.337. **G)** Mutual information analysis for the association of syllable use with estrous phase, individual identity and the interaction of phase and identity. Each datapoint represents a single syllable. Top 10 syllables associated with differences between phases included: runs and darts, running left/right, groom, rears and pause. Syllables associated with differences between individuals include: rears against the arena wall, darts, rears, and turns, although every syllable had greater mutual information with individual identity than information about estrous phase. **H)** Classification accuracy of decoders trained to predict phase and mouse identity based on syllable bigrams (sequences of two syllables, see Methods). For all relevant panels, Box plots depict median, interquartile range, and upper/lower adjacent values (black lines). For all decoder panels: asterisk (*)

denotes statistical significance, here indicating that the mean decoding distribution exceeds the 95th percentile shuffle threshold. For all other panels: *p<0.05, **p<0.01, ***p<0.001

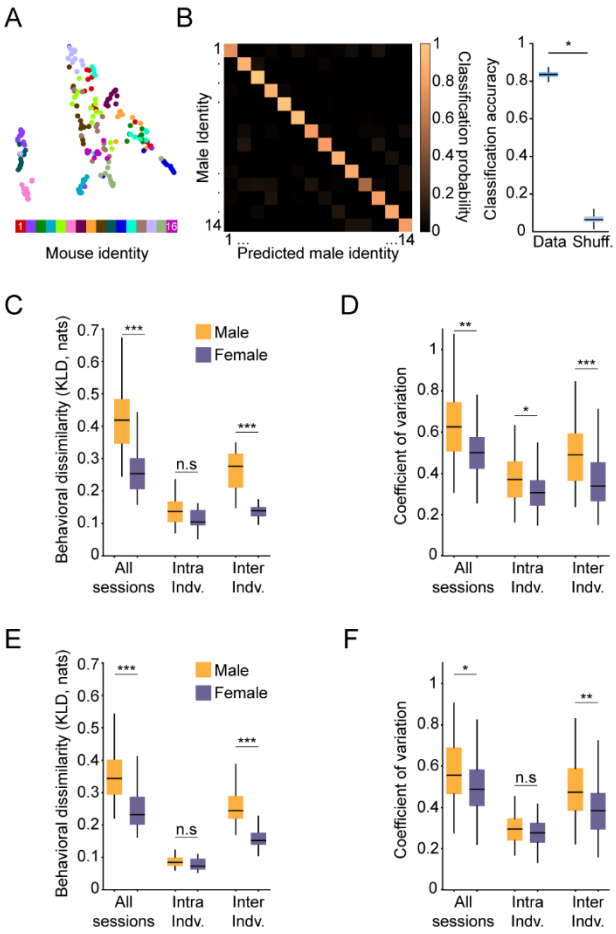


Figure S3: Male behavior is more variable than female behavior in the open field. related to Figure 4.

A) Uniform Manifold Approximation and Projection (UMAP) plot depicting syllable usage in males for each session, colored by mouse identity. To assess cluster quality, K-means clustering analysis was performed on high-dimensional data and clustering quality compared to true labels was quantified using the Adjusted Rand Index (ARI). For individuals (number of clusters = 16) ARI = 0.47. **B)** Classification accuracy for male individual identity based on syllable usages. Left: confusion matrix for decoder performance. Right: overall decoder performance is presented against shuffled data (Shuff.). Asterisk (*) denotes statistical significance, here indicating that the mean decoding distribution exceeds the 95th percentile shuffle threshold. **C)** Comparison of behavioral variability as measured by the KLD of syllable usage distributions during the first week of female behavioral sessions and the last week of male behavioral sessions. Pairwise comparisons were done between syllable usage distributions in each behavioral sessions (“all sessions”), between sessions within each individual (intra indv. = intra-individual variability), and between individuals (inter indv. = inter-individual variability). All comparisons were done between individuals from the same sex. 2-way ANOVA for sex and experimental conditions (exp) as main factors was performed:

$Fsex_{(1,249)}$=34.8, p=1.18*10^{-8}, $Fexp_{(2,249)}$=18.41, p=3.47*10^{-8}, $Fsex*exp_{(2,249)}$=1.95, p=0.144. Individual contrasts were performed using student's t-test with Bonferroni correction: $p_{all\ sessions}$=1.09*10^{-6}, n_{female}=93 sessions, n_{male}=98 sessions; $p_{intra\text{-}indv}$=1; $p_{inter\text{-}indv}$=2.9*10^{-6}, n=16 mice for male and females. **D**) Same as in C. but for CV of syllable usage, calculated per syllable. $Fsex_{(1,288)}$=31.7, p=4.18*10^{-8}, $Fexp_{(2,288)}$=56.46, p=2.04*10^{-21}, $Fsex*exp_{(2,288)}$=1.75, p=0.176. Individual contrasts were performed using student's t-test with Bonferroni correction: $p_{all\ sessions}$=0.0027; $p_{intra\text{-}indv}$=0.036; $p_{inter\text{-}indv}$=0.0009. n_{mice}=16 and $n_{syllables}$=49 for both male and females. **E)** Same as C, but for an independent dataset, comparing male and female behavioral variability when matching handling conditions and following prolonged (10 days) habituation to the experimental setup. $Fsex_{(1,252)}$=88.57, p=3.2*10^{-18}; $Fexp_{(2,252)}$=128.39, p=3.56*10^{-39}, $Fsex*exp_{(2,252)}$=4.77, p=0.009. Individual contrasts were performed using student's t-test with Bonferroni correction: $p_{all\ sessions}$=1.14*10^{-14}, n_{female}=94 sessions, n_{male}=100 sessions; $p_{intra\text{-}indv}$=0.51; $p_{inter\text{-}indv}$=0.0008, n=16 mice for male and females. **F)** Same as in E, but measured as CV of syllable usage, per syllable. For left panel: $Fsex_{(1,288)}$=17.97, p=3.02*10^{-5}; $Fexp_{(2,288)}$=102.83, p=1.9*10^{-34}, $Fsex*exp_{(2,288)}$=2.37, p=0.09. Individual contrasts were performed using student's t-test with Bonferroni correction: $p_{all\ sessions}$=0.036; $p_{intra\text{-}indv}$=0.63; $p_{inter\text{-}indv}$=0.0051, n_{mice}=16 and $n_{syllables}$=49 for both male and females. For all relevant panels, Box plots depict median, interquartile range, and upper/lower adjacent values (black lines). *p<0.05, **p<0.01, ***p<0.001

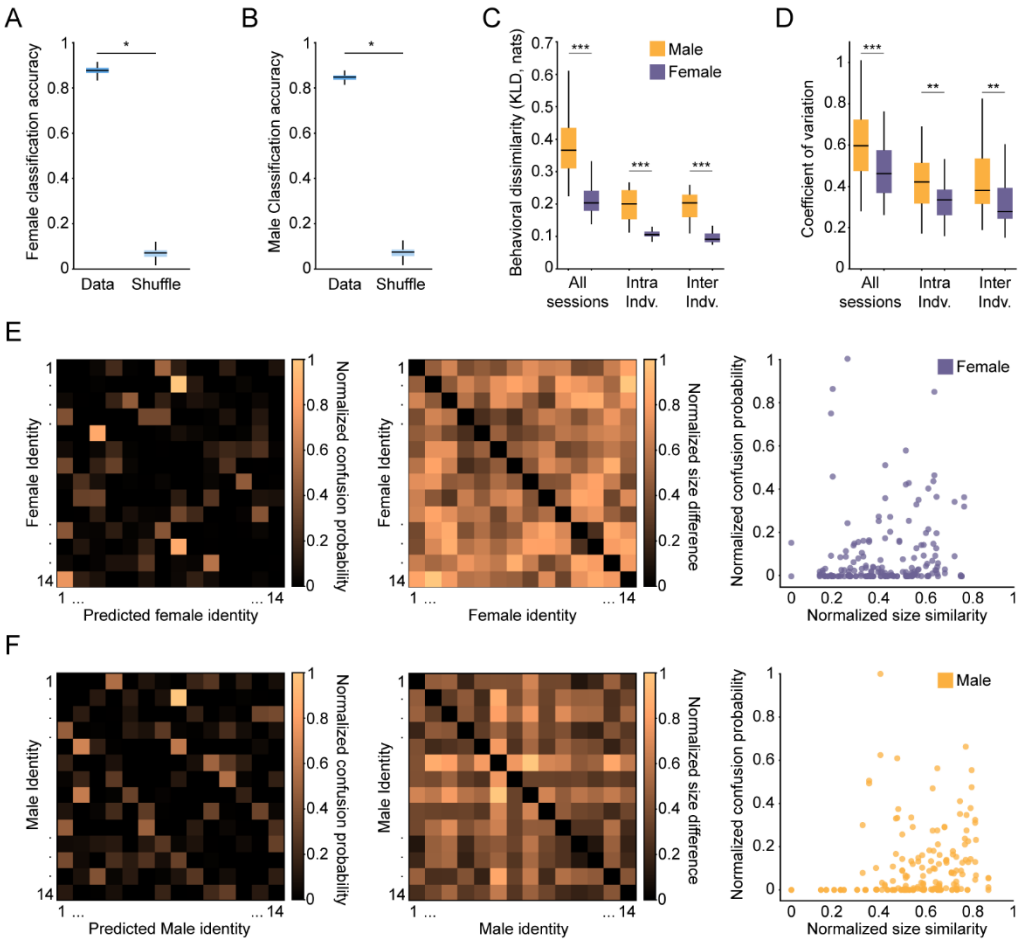


Figure S4: Size differences do not account for inter-individual behavioral variability as measured by MoSeq, related to Figure 4.

A) Classification accuracy for female individual identity based on syllable usages, after exclusion of syllable whose usage correlates with mouse size (see Methods). Asterisk (*) denotes statistical significance, here indicating that the mean decoding distribution exceeds the 95th percentile shuffle threshold. **B)** Same as A but for male individual identity. **C-D)** Behavioral variability analysis as in Fig. 2C,D but after exclusion of syllables whose usage correlates with mouse size (see Methods). **C)** Comparison of male and female behavioral variability between all behavioral sessions, between sessions within each individual (intra indv. = intra-individual variability), and between individuals (inter indv. = inter-individual variability) as measured by KLD. All comparisons were done using individuals from the same sex. 2-way ANOVA for sex and experimental conditions (exp) as main factors was performed: $Fsex_{(1,438)}=87.1$, $p=5.2*10^{-19}$; $Fexp_{(2,438)}=16.19$, $p=1.64*10^{-7}$, $Fsex*exp_{(2,438)}=2.14$, $p=0.117$. Individual contrasts were performed using student's t-test with Bonferroni correction: $p_{all\ sessions}=1.35*10^{-15}$, $n_{female}=188$ sessions, $n_{male}=192$ sessions; $p_{intra-indv}=6.87*10^{-6}$, n=16 mice for male and females; $p_{inter-indv}=9.93*10^{-7}$, n=16 mice for male and females. **D)**

Same as in C, but depicting the distribution of coefficients of variation of the usage of each syllable. For left panel: 2-way ANOVA: $Fsex_{(1,219)}=36.33$, $p=6.96*10^{-9}$; $Fexp_{(2,219)}=33.42$, $p=2.15*10^{-13}$, $Fsex*exp_{(2,219)}=0.55$, $p=0.57$. Individual contrasts were performed using student's t-test with Bonferroni correction: $p_{all\ sessions}=0.0009$; $p_{intra-indv}=0.0024$; $p_{inter-indv}=0.0015$. $n_{syllables}=41$ for males and $n_{syllables}=34$ for females. **E)** Left: Confusion matrix depicting the output of a "held-out" classifier, which enables classifier-based quantification of the similarity of behavioral patterns expressed by pairs of female mice (see Methods). Middle: Euclidian distances between size measurements of individual females (see Methods). Size differences and classification probabilities, depicted by colormap, were normalized to 0-1 scale. Right: Size similarity (defined as "1-x", x=size difference) plotted against decoder confusion probabilities. For linear fit, R^2 = 0.014. Self-distances were removed from the analysis. **F)** Same as E, but for males. For linear fit, R^2 =0.03. For all relevant panels, Box plots depict median, interquartile range, and upper/lower adjacent values (black lines). *p<0.05, **p<0.01, ***p<0.001

Syllable ID	Associated behavior	MI score phase	MI score identity	CV between individuals
0	Rear	0	0.29	0.24
1	Dart	0	0.57	0.24
2	Low rear	0	0.32	0.36
3	Groom	0.01	0.32	0.19
4	Down from wall rear	0	0.2	0.16
5	Run left	0.04	0.44	0.26
6	Groom	0	0.21	0.15
7	Pause before rear	0	0.35	0.24
8	Turn left	0	0.61	0.27
9	Walk	0.03	0.37	0.34
10	Run	0.03	0.2	0.27
11	Short dart	0.01	0.3	0.24
12	Walk	0.06	0.16	0.25
13	Groom	0.02	0.45	0.22
14	Groom	0.0	0.07	0.19
15	Short dart	0.1	0.38	0.33
16	Dart	0.07	0.35	0.49
17	Down from rear	0.02	0.37	0.34
18	Run right	0.0	0.35	0.27
19	Groom	0.09	0.25	0.38
20	Rear	0.0	0.31	0.24
21	Rear	0.02	0.25	0.40
22	Walk	0.0	0.14	0.26
23	Low rear	0.0	0.2	0.26
24	Down from wall rear	0.03	0.4	0.24
25	Dart	0.0	0.22	0.20
26	Rear	0.0	0.33	0.27
27	Scrunch	0.0	0.38	0.52
28	Wall rear	0.0	0.77	0.52
29	Run	0.04	0.17	0.22
30	Wall rear	0.0	0.64	0.48
31	Wall rear	0.02	0.24	0.25
32	Run right	0.05	0.21	0.26
33	Short pause	0.0	0.31	0.36
34	Scrunch	0.0	0.28	0.24
35	Groom	0.01	0.2	0.31
36	Rear	0.03	0.43	0.56
37	High rear	0.0	0.25	0.28
38	Stretch	0.0	0.25	0.63
39	Rear	0.0	0.28	0.40
40	Wall rear	0.0	0.5	0.45
41	Pause	0.03	0.12	0.16
42	Run left	0.0	0.47	0.60
43	Groom	0.0	0.29	0.41
44	High rear	0.0	0.29	0.35
45	Low rear	0.0	0.23	0.45
46	Rear	0.04	0.55	0.72
47	Low rear	0.0	0.32	0.72
48	High rear	0.02	0.26	0.49

Table S1: Syllable labels for females, related to Figures 2,4 and S2.

This table lists all of the behavioral syllable identified in the MoSeq model used for female mice, the human-annotated behaviors associated with each syllable, as well as the degree of mutual information for estrous phase and individual identity (see Methods); also shown is the coefficient of variation of the usage of each syllable across individuals.

Syllable ID	Associated behavior	MI score identity	CV between individuals
0	Pause	0.53	0.22
1	Low rear	0.51	0.38
2	Stretch	0.52	0.36
3	Dart	0.49	0.37
4	Short rear	0.55	0.46
5	Down from rear	0.56	0.37
6	Wall rear	0.27	0.19
7	Walking	0.27	0.30
8	Pause	0.08	0.26
9	Run	0.45	0.42
10	Turn left	0.39	0.28
11	Short pause	0.22	0.32
12	Short dart	0.12	0.18
13	Groom	0.77	0.57
14	Run right	0.53	0.45
15	Low rear	0.54	0.50
16	Run	0.63	0.61
17	High rear	0.25	0.28
18	Down from rear	0.34	0.39
19	Pause	0.19	0.44
20	Dart	0.24	0.36
21	Turn right	0.75	0.42
22	Walk	0.23	0.31
23	Rear	0.37	0.32
24	Short dart	0.44	0.53
25	Rear	0.22	0.27
26	High rear	0.38	0.53
27	High rear	0.42	0.55
28	Down from wall rear	0.31	0.25
29	Run right	0.28	0.41
30	Groom	0.38	0.36
31	High rear	0.47	0.53
32	Low rear	0.48	0.67
33	Rear	0.56	0.55
34	High rear	0.46	0.77
35	Rear	0.36	0.51
36	Run left	0.39	0.37
37	Pause	0.52	0.50
38	Scrunch	0.85	0.82
39	Run right	0.45	0.42
40	Groom	0.24	0.50
41	Run	0.37	0.56
42	Groom	0.21	0.37
43	Groom	0.23	0.35
44	Wall rear	0.33	0.29
45	Stretch	0.59	0.95
46	Stretch	0.29	0.56
47	High rear	0.4	0.71
48	Rear	0.33	0.44

Table S2: Syllable labels for males (related to Figure 4)

This table lists all of the behavioral syllable identified in the MoSeq model used for male mice, the human-annotated behaviors associated with each syllable, as well as the degree of mutual information for individual identity (see Methods); also shown is the coefficient of variation of the usage of each syllable across individuals.