Real-time analysis of the behaviour of groups of mice via a depth-sensing camera and machine learning

Fabrice de Chaumont1,6*, Elodie Ey2,6*, Nicolas Torquet3, Thibault Lagache1, Stéphane Dallongeville1, Albane Imbert4, Thierry Legou5, Anne-Marie Le Sourd2, Philippe Faure3,7, Thomas Bourgeron2,7 and Jean-Christophe Olivo-Marin1,7*

Preclinical studies of psychiatric disorders use animal models to investigate the impact of environmental factors or genetic mutations on complex traits such as decision-making and social interactions. Here, we introduce a method for the real-time analysis of the behaviour of mice housed in groups of up to four over several days and in enriched environments. The method combines computer vision through a depth-sensing infrared camera, machine learning for animal and posture identification, and radio-frequency identification to monitor the quality of mouse tracking. It tracks multiple mice accurately, extracts a list of behavioural traits of both individuals and the groups of mice, and provides a phenotypic profile for each animal. We used the method to study the impact of Shank2 and Shank3 gene mutations—mutations that are associated with autism—on mouse behaviour. Characterization and integration of data from the behavioural profiles of Shank2 and Shank3 mutant female mice revealed their distinctive activity levels and involvement in complex social interactions.

Mice are routinely used as preclinical models to study the mechanisms of human diseases. In psychiatry, assessing the social behaviour of mice under normal or pathological conditions is critical for the understanding of which neural systems are engaged in disease. However, such behaviours are complex and challenging to investigate in mice, given the technical limitations of data gathering and analysis. Indeed, although behavioural protocols that investigate one or two animals can provide information on activity and cognitive functions, such as learning, memory and anxiety, gathering appropriate information on social behaviours requires access to the observation of more than two individuals. Highly standardized social interaction tests are available but these tests rely primarily on the quantification of simple (dyadic) and short (a few minutes) social interactions (which are often manually labelled), and lack ethologically relevant behavioural markers (such as the maintenance of social interactions over long periods of time and social interactions that involve more than two mice). Furthermore, it is necessary to increase the robustness of the data collected and to provide additional extracted features to thoroughly document the dynamics and plasticity of the social aspects over unlimited periods of time for each individual. This is particularly true, for example, for the formation and description of group dynamics of animals.

Solutions to automatically track mice have been proposed (see the ‘Motivation and review of the existing tracking methods’ section in the Supplementary Information). They are all technically constrained and therefore are dedicated, more or less, to a specific phenotypic extraction type. By choosing a system, an experimenter is then bound to its set of constraints, which include the tracking reliability, the amount of manual corrections, the potential need to visually mark animals, the potential need to implant a radio-frequency identification (RFID) probe into the animal, the possibility of accessing or building the system, the ease of use of the software and the ability to obtain exploitable data in a reasonable time. Owing to this extensive range of parameters—and despite the number of existing tracking systems—the choice is limited in practice, as data extracted vary from highly detailed segmentation at high frame rates to rough localization information at low frame rates. With increasing recording durations, all systems progressively provide fewer details in the data extracted. No existing system (whether commercial or academic) provides all these resources at once. Therefore, researchers have to choose and trade some demands in exchange for others—for example, the choice between a long duration of observation and the depth of data information. Our flexible and open source system was designed to provide unlimited observation duration, a high accuracy of tracking and the possibility of interacting with the assay in real time under the control of computed track parameters.

Our comprehensive system, which we call Live Mouse Tracker (LMT), enables automatic live tracking, identification and characterization through behavioural labelling of up to four mice in an enriched environment with no time limit. This solution makes use of a random-forest machine learning process, RFID sensors and an infrared–depth RGB-D camera. The infrared–depth camera is used to distinguish mice from the background; the system therefore works in three dimensions and can detect mice regardless of their appearance. Once mice are segmented, the random-forest machine learning process enables their orientation to be determined, as well as...
as their identity. Identities are constantly validated using RFID tags detected by floor antennas that are selectively activated when an unidentified mouse passes over an antenna. Behavioural labelling is performed in real time using the continuous stream of depth and volume information to characterize the shape and posture of each individual (see a previously published example of tracking a single mouse19), as well as the relative positions between individuals.

On the basis of this robust central element, we designed a system that can address the whole behavioural assay workflow and cope with most—if not all—of the difficulties of this type of assay: (1) reproducible acquisition hardware (Supplementary Video 2; see the ‘Assembly instructions’ section in the Supplementary Information); (2) acquisition calibration (see the ‘Camera setup and calibration’ section in the Methods); (3) standardized data acquisition; (4) live tracking (identity recovery and identity control; see the ‘Triggered-RFID closed-loop control for identification’ section in the Methods); (5) automatic phenotyping (see the ‘Behavioural event extraction’ section in the Supplementary Information); (6) on-the-fly analysis for monitoring the tracking quality or the behaviour (see the ‘Automatic tracking control’ section in the Supplementary Information); (7) storage of long-term observations in a database (see the ‘Database description’ section in the Supplementary Information); (8) visual data inspection with the development of a specific data player (see the ‘Database player’ section in the Supplementary Information); and (9) online data sharing with the community (databases, videos, analysis scripts and results; see the ‘Collaborative website sharing data’ section in the Supplementary Information).

We used LMT with up to four animals at a time to compare the behavioural profile of female mice that lack Shank2−/− or Shank3−/− two genes that were found to be mutated in a subgroup of patients with autism and that encode post-synaptic scaffolding proteins in excitatory synapses10,11. We report behavioural differences in the activity levels and involvement in complex social configuration between the individual profiles of the Shank2 and Shank3 strains. We confirm that the characteristic differences in activity levels between Shank2−/− and Shank3−/− mice were still present, even after long periods of interaction. Despite these different activity levels, Shank2−/− and Shank3−/− mice displayed typical circadian rhythms. Finally, we show that the atypical social behaviour of Shank2−/− and Shank3−/− mice appeared to disturb the formation of subgroups within mixed-genotype groups of four mice.

These results demonstrate the capability of LMT to study differences in the phenotypes expressed by individuals in groups of mice over long time periods.

Results

We designed an integrated system that tracks and monitors—for hours or days—the activities of four mice in a rich environment (Supplementary Video 1). This system determines the outline mask and orientation of each mouse and builds a comprehensive repertoire of individual and social behavioural events from these data. The system uses machine learning that enables the tracking of mice of any coat colour (black, white or agouti; Supplementary Video 7). The machine learning implementation provides a real-time discrimination between mice and food, water, sawdust, Video 7). The machine learning implementation provides a real-time discrimination between mice and food, water, sawdust, activity levels and involvement in complex social configuration (Supplementary Information) and can be investigated live during tracking (see the ‘Database player’ section in the Supplementary Information). Video and background maps are simultaneously recorded as videos and image series, respectively (see the ‘MP4 video recording’ section in the Supplementary Information).

Tracking reliability was assessed in four experiments, each with a duration of 10 min (that is, 18,000 frames for each experiment) with one, two, three or four mice in the same cage. LMT detected the mice at least 99.25% of the time (that is, the detection rate), for any number of mice (from one to four; Fig. 1). Two independent experts then validated the detection of the mice (see the ‘Manual validation (extra information)’ section in the Supplementary Information; Supplementary Figs. 3–5) to estimate the reliability of (1) segmentation (Fig. 1k), which is essential to determine whether mice are in contact and to conduct shape analysis; (2) head–tail orientation (Fig. 1l), which is needed for asymmetrical events and (3) identification of individual mice (Fig. 1m), which is essential to understand interindividual relationships. The experts used the integrated database player (see the ‘Database player’ section in the Supplementary Information) to manually validate the tracker performance frame by frame. In the group of four, mice were correctly segmented (that is, the mask fit the real body shape exactly, extra objects or reflections on the walls were not included and animals were not merged) in more than 95.75% of the frames in which a...
mouse was detected (Fig. 1k). The detected orientation was accurate in more than 99.36% of the detected frames (Fig. 1l). Finally, the identity error rate did not exceed 2.69% for a group of four mice (Fig. 1m). Overall, the system keeps track of the identities of the mice, corrects misidentifications and prevents any error from propagating thanks to the RFID used in the system. Episodes in which identities were switched had a mean duration of 1.64 s ± 0.23 s for expert 1, and 1.33 s ± 0.23 s for expert 2 (see the ‘Manual validation (extra information)’ section in Supplementary Information; Supplementary Fig. 3a). All of these estimates allowed us to calculate the multiple object tracking accuracy (MOTA) tracking performance index13, which considers false-positive and false-negative identifications, and identity switches. The MOTA reached 0.993, 0.991, 0.984 and 0.970 using one to four mice, respectively (see the ‘Manual validation (extra information)’ section in the Supplementary Information; Supplementary Fig. 3b). As a comparison, a tracking system for pig behaviour reached a maximum of 0.90 for the MOTA index14.

Finally, it should be noted that beyond this manual validation, the system checks the mouse identities by constantly comparing the IDs stored by the tracker with those detected by RFID. The system then logs the number of either ID confirmations (match) or ID corrections (mismatch) in the database. Users can therefore monitor the tracking quality of their experiments using the metrics described in the Supplementary Information or metrics of their own choosing (see the ‘Automatic tracking control’, ‘Automatic tracking control based on RFID’ and ‘Automatic tracking control based on detection’ sections in the Supplementary Information).

**Fig. 1 | Tracking method and validation.** a, A cage encloses four black mice in an enriched environment that contains food, water, sawdust, brown crinkle paper, white cotton cylinders, toys and a red house (opaque to mice, transparent to infrared). b, The software display (infrared view) that corresponds to the image in a. Each mouse is represented, including its current segmentation. Coloured lines represent tracks during the last 60 frames. Time codes (experimental and real) are displayed at the top. Circle magnifications on each side represent mice individually, reoriented to face upwards. For each magnified image, the following information is displayed: the RFID number of the mouse, the posture names (rearing, looking up, looking down; white if active, otherwise black; for example, the yellow mouse is rearing and looking up) and the Z-profile (plot of the height of the animal along its longitudinal axis) of the main axis of the animal with optional detection of ears (red) and nose (green). c–i, Example segmentations of white mice, showing the acquired infrared image (c,d), corresponding depth map (e), subtraction of the acquired depth map and background map (f), three-dimensional representation of the detection of the mice (colour scale indicates height from 0 (red) to 3 (green) cm) (g) and a close-up view of the head of a mouse (scale bar, 1 cm) (h) with detected nose and ears (i). j–m, Manual validation performed by two experts over 10 min on 1 to 4 mice. j, Detection error rate—the proportion of frames in which the animal is not detected (the percentage of 18,000 frames). k, Segmentation error rate—the proportion of frames in which the animal is detected and well segmented but the head and tail are not correctly detected or reversed. l, Orientation error rate—the proportion of frames in which the animal is detected but not correctly segmented (that is, its shape is not exact). m, Identity error rate—the proportion of frames in which the animal is correctly detected and segmented but its identity is not correct.
For each time frame and each detected mouse, LMT provides the mask of the animal, its depth mask relative to the background, the location of the head and tail, and detection of the ears, eyes and nose. This data served as the basis for computing a number of events (see the 'Behavioural event extraction' section in the Supplementary Information) that were inferred from shape geometry. Overall, we defined 35 behavioural events related to intrinsic and relative positions of the mice that were extracted automatically. The events can be split into five categories (Fig. 2a; see the 'Behavioural event extraction' section in the Supplementary Information): (1) events related to individual behaviour that reflect activity and individual postures, (2) social dyadic events that reflect the different types of contact between two mice (for an example of the different types of contact established between pairs of mice during 4 min, see the 'Social contact timelines' section in the Supplementary Information; Supplementary Fig. 6), (3) dynamic...
events involving the approach, escape and following behaviours involving two mice, (4) configuration events that reveal subgroup configurations with two, three or four mice, and (5) group-making and group-breaking events that focus on the dynamics leading to the creation or the ending of subgroups. All of these behaviours were computed during long-term experiments (Fig. 2b, chronogram). We manually validated the different types of contacts (general contact, nose-to-nose contact, nose-to-anogenital contact and side-by-side contact), because all social events were based on these contacts and on geometric formulae (see the ‘Manual validation (extra information)’ section in the Supplementary Information; Supplementary Figs. 3–5).

**Individual profiles of Shank2−/− and Shank3−/− mice.** To demonstrate the potential of the LMT, we compared the behavioural profiles of two mouse models of autism spectrum disorder with mutations in genes—Shank2 and Shank3—that code for two synaptic scaffolding proteins of the Shank family. Despite the relatedness of the two proteins, subtle behavioural differences between these two models were expected due to the differences in expression profiles of these two proteins.

We tracked four mice individually within well-established groups to quantify the exploration of the environment and the social interactions within a test cage with fresh bedding and enrichment material. We designed a configuration—based on two wild-type and two mutant mice—that allowed the simultaneous monitoring of the control relationship between wild type and wild type, the interaction between wild type and mutant and the interaction between mutant and mutant to provide all possible dyadic relationships that are normally tested using two animals in classical settings. We recorded nine mixed-genotype groups of four mice for the Shank2 strain over 23 h and six mixed-genotype groups of four mice for the Shank3 strain over 3 d. Unless otherwise specified in the text, we analysed only the first 23 h of the Shank3 recordings to retain comparability with Shank2 mice.

In this analysis, LMT automatically extracted 35 behavioural traits for each individual mouse (see the ‘Behavioural event extraction’ section in the Supplementary Information); these were the traits that could be compared between wild-type and mutant mice. The distribution of a subsample of behavioural traits (Supplementary Fig. 8) revealed that the behaviour of Shank2−/− mice was more affected by the mutation than the behaviour of Shank3−/− mice. Shank2−/− mice moved significantly more and spent significantly shorter time in side-by-side contact with individuals of the same genotype compared with wild-type mice (Supplementary Fig. 8). None of the behavioural traits examined—including social events—were significantly affected in Shank3−/− mice, confirming that Shank3−/− mice displayed only subtle behavioural abnormalities (see Supplementary Table 3). To compare mutant and wild-type mice with the same baseline and to avoid interexperiment variability, the value of each trait (in duration or in number of events) for one mutant was compared with the mean level of this trait for the two wild-type mice of their respective cage. We could therefore detect whether the specific behavioural trait increased or decreased in mutant mice compared with the mean level of wild-type mice tested under exactly the same conditions (that is, within the same cage).

The behavioural profile of Shank2−/− mice (Fig. 3a; see Supplementary Fig. 9 for individual data and Supplementary Table 4 for statistical analyses) reflected higher locomotion activity (time spent moving alone, moving in contact and stopped alone) and lower exploration (duration of stretched attend posture (SAP) for individual events. In social events, Shank2−/− mice displayed reduced time spent in side-by-side contact (same and opposite way), slower social approaches that lead to a contact (make contact duration), increased following behaviour (duration of a train of two and duration of follow), and increased frequency of completing or breaking a group of three or four mice (make a group of three, break a group of three and make a group of four). By contrast, the behavioural profile of Shank3−/− mice (Fig. 3b; for individual data, see Supplementary Fig. 10) indicated significantly reduced activity (moving alone) and significantly reduced time spent in a complex social interactions that involved three mice (that is, time spent at the end of a long line of three mice (train of three)). We next compared the behavioural profiles of the Shank2−/− mice and the Shank3−/− mice to identify the subtle effects of these mutations. Significant differences emerged between the Shank2 and Shank3 profiles in activity measures (Wilcoxon rank sum tests with Bonferroni corrections for multiple testing: time spent moving alone (W = 213, P = 0.00003) and in contact (W = 214, P < 0.00003), and time spent stopped alone (W = 14, P = 0.00003)), in exploratory measures (time spent in SAP (W = 29, P = 0.00044)), in time spent in side-by-side contact (opposite (W = 28, P = 0.00036) and same way (W = 32, P = 0.00079)), in the duration of social approaches (duration of making contact (W = 190, P = 0.00024)) and in the dynamic of groups of three (making (W = 200, P < 0.00003) and breaking groups of three (W = 196, P = 0.00006)). Overall, despite carrying mutations in genes from the same family, these two mouse models of autism spectrum disorder displayed inverse phenotypes regarding activity and exploration, as well as different alterations in social behaviours. These phenotypic variations remained when the profile of Shank2−/− mice was compared with the profile of 10–11-month-old Shank3−/− mice (Supplementary Fig. 11; for statistical analyses, see Supplementary Table 4).

We then investigated the specific aspects of the differences between the two mouse models in further detail, first focusing on specific social behaviours. We conducted these investigations during the same experiments by computing the data extracted from the long-term monitoring of a group of four mice.

**Exploring the subgroup dynamics.** We expected mouse models of autism spectrum disorder to exhibit deficits in social features, such as more frequently remaining isolated from the rest of the group compared with wild-type mice. We thus provide a comprehensive analysis of the motivation of each mouse to join or leave a particular social structure. The social structure is defined as the periods when the mice form groups, that is, when they are in contact or close proximity with one another (Fig. 2a).

We focused on groups of three mice, and first noticed that the two types of groups of three mice (two wild-type and one mutant, or two mutant and one wild-type) were equally frequent. This suggested that—contrary to our expectations—Shank2−/− or Shank3−/− mice were not more frequently isolated than their wild-type littermates when a group of three mice was formed within the cage (Supplementary Fig. 12a,b). This also revealed that the hyperactivity of Shank2−/− mice did not seem to perturb their ability to form a group of three mice.

We then examined group dynamics to assess whether mutant mice were less attracted by social interactions involving more than one mouse. To this end, we first determined which individual completed or broke the group in groups of three mice (see the Dynamics of subgroups of three mice within groups of four mice section in the Supplementary Information). We defined the chance level as the probability of an individual with a given genotype joining or leaving a group of mice divided by the total number of possible combinations. With two wild-type and two mutant mice within the cage, there were 12 possible combinations of groups of three mice when considering the initial members and the joiner or breaker (that is, ordered combination; Fig. 4). Variations in probability compared with chance levels indicated that one genotype was more (in the case of a higher probability than chance level) or less (in the case of a lower probability than chance level) likely than the other genotype to join or to leave a group. For both Shank2 and Shank3 strains, the probability of a knockout mouse joining/leaving a pair...
of wild-type mice, or the probability of a wild-type mouse joining/leaving a pair of knockout mice was higher than expected by chance (joining: one-sample Student’s t-tests, exact P-values are marked with an asterisk when significance remains after Bonferroni correction for multiple testing: where initial members were wild type (+/+) and +/−, and the joining member was knockout (−/−)); Shank2: t = 23.324, d.f. = 8, *P = 4.214 × 10⁻⁴; Fig. 4a; Shank3: t = 22.564, d.f. = 5, *P = 3.178 × 10⁻⁴; Fig. 4c; initial members −/− and −/−, joining member +/+: Shank2: t = 14.45, d.f. = 8, *P = 5.147 × 10⁻⁵; Fig. 4a; Shank3: t = 19.574, d.f. = 5, *P = 6.423 × 10⁻⁴; Fig. 4c; breaking: t-tests: initial members +/+ and +/+, breaking member −/−: Shank2: t = 28.342, d.f. = 8, *P = 2.596 × 10⁻⁴; Fig. 4b; Shank3: t = 28.704, d.f. = 5, *P = 9.616 × 10⁻⁴; Fig. 4d; initial members −/− and −/−, breaking member +/+: Shank2: t = 21.708, d.f. = 8, *P = 2.138 × 10⁻⁴; Fig. 4b; Shank3: t = 23.566, d.f. = 5, *P = 2.561 × 10⁻⁴; Fig. 4d). Therefore, in both models, pairs of mice with the same genotype (either mutant or wild type) seem to be more attractive and more repulsive to the other genotype than expected by chance. As Shank2 and Shank3 strains were similar despite their difference in activity levels, this mutant–wild type distinction could not be explained by hyperactivity or hypoactivity. Interestingly, in both strains, the mean duration of a group of three event (−/−, −/− and +/+, or +/−, +/− and −/−) created by a mouse of a given genotype and broken by a mouse of the same genotype tended to be shorter than in groups in which the joiner and the breaker were of different genotypes (Wilcoxon signed-rank tests with Bonferroni correction for multiple testing: Shank2: P < 0.05; Shank3: P < 0.1; Supplementary Fig. 13). This suggested that the duration of groups of three that were created and broken by the same individuals were shorter than those created and broken by different individuals. These short group of three events also reflected the fact that mice have a tendency to pass near a group of two without stopping.

We also explored the completion and breaking of groups of four mice. The probability of joining a group of four last or to leave the group first was not significantly different between wild-type and mutant mice of both strains (Fig. 4e–h). This suggested that the level of activity did not disturb the social grouping ability of Shank2/−/− or Shank3/−/− mice when all mice in the cage are involved.

Together, these results indicate that our methodology allows us to assess the dynamics of group formation, behavioural information that—owing to the technical limitations of current techniques—has rarely been described (except in ref. 17). Such possibilities are of high interest for mouse models such as Shank2/−/− and Shank3/−/− mice that display a large variability in the severity of their social phenotype according to their genetic construction and to experimental conditions (reviewed previously for Shank2/−/− mice18 and Shank3/−/− mice19).

**Activity levels in groups.** The comparison of the individual profiles between Shank2/−/− mice Shank3/−/− mice revealed opposite levels of activity (see above), confirming results obtained previously using classical protocols17 (Supplementary Table 3). By following the group-housed mice individually with our system, we were also able to assess how these differences in activity level affected the circadian activity of the mice. Over 23 h of monitoring, Shank2/−/− mice travelled significantly longer distances (5,107 ± 300 m) than their wild-type littermates (3,121 ± 134 m; Wilcoxon rank-sum test: W = 12, P < 0.001). Shank2/−/− mice first displayed a novelty-induced hyperactivity at the beginning of the recording. However, despite their strong hyperactivity, Shank2/−/− mice exhibited typical circadian rhythms, as suggested in another model of Shank2/−/− mice in which exon 24 was deleted15. Indeed, active and resting periods in Shank2/−/− and Shank2/−/− mice were synchronous and the general locomotor activity over 23 h was significantly correlated between Shank2/−/− and Shank2/−/− mice ( Spearman rank correlation: ρ = 0.951, S = 2,660, P < 0.001; Supplementary Fig. 14). Shank3/−/− mice tended to travel shorter distances over the first day (2,970 ± 75 m) compared with wild-type littermates (3,287 ± 116 m; Wilcoxon rank-sum test: W = 101, P = 0.101). The activities
of Shank3\(^{+/−}\) and Shank3\(^{−/−}\) mice were significantly correlated (Spearman rank correlation: \(ρ = 0.958, S = 2.292, P < 0.001\)) and did not suggest that Shank3\(^{−/−}\) mice were hypoactive specifically because of the novelty of the environment, but they were constitutively hypoactive. We prolonged the monitoring of these mice over 3 d to confirm this hypoactivity. Over 3 d, Shank3\(^{−/−}\) mice travelled significantly shorter distances (7,631 ± 142 m) compared with Shank3\(^{+/−}\) mice (8,374 ± 226 m; Wilcoxon rank-sum test: \(W = 112, P = 0.020\)). Again, the active and resting periods were synchronous between Shank3\(^{+/−}\) and Shank3\(^{−/−}\) mice (Spearman rank correlation: \(ρ = 0.964, S = 61.024, P < 0.001\); Supplementary Fig. 14). Overall our data revealed a novelty-induced hyperactivity and a general hyperactivity that was embedded in the classical circadian rhythms of Shank2\(^{−/−}\) mice. This hyperactivity, measured by the computation of the total distance travelled, confirmed the increased time spent moving (either alone or in contact with another mouse) and reduced time spent stopped alone identified in the individual profile (Fig. 3a). In Shank3\(^{−/−}\) mice, we observed a general hypoactivity that was embedded in the classical circadian rhythms, which also confirmed the reduced time spent moving alone in the individual profile (Fig. 3b). Interestingly, the locomotor hyperactivity of Shank2\(^{−/−}\) mice might perturb exploratory behaviour according to the reduction of SAP behaviour (Fig. 3a), whereas locomotor hypoactivity did not seem to have any effect on exploratory behaviour (Fig. 3b). Therefore, we next addressed the dynamic of the hyperactivity and hypoactivity using a new environment and the interaction with exploratory behaviours.

**Activity and object exploration in single and dyadic tasks.** Our system was flexible enough to investigate the differences in activity levels in single (Shank2\(^{−/−}\) and Shank3\(^{−/−}\) mice) and dyadic (Shank2\(^{−/−}\) mice) exploration tasks. During the 30 min habituation in the test cage filled with fresh bedding (phase 1; Fig. 5a), Shank2\(^{−/−}\) mice travelled significantly longer distances compared with Shank2\(^{+/−}\) littermates (Wilcoxon rank-sum test: single: \(W = 0, P < 0.001\); paired: \(W = 0, P < 0.001\); Fig. 5b,c, phase 1). These data corroborate results that were obtained in the same animals using the classical open-field test (Spearman correlation between the open field test and the single exploration test: \(ρ = 0.727, P < 0.001\); data not shown). The level of activity measured in pairs was significantly correlated with the level of activity measured during single exploration of the test cage (Spearman correlation: phase 1; \(ρ = 0.899, P < 0.001\)). By contrast, Shank3\(^{−/−}\) mice travelled significantly shorter distances compared with their wild-type littermates in phase 1 (Wilcoxon rank-sum test: single: \(W = 81, P < 0.001\), confirming their hypoactivity in a new environment.

Finally, we investigated the influence of the activity level on object exploration. For this purpose, we introduced on-the-fly (that...
is, without stopping experiment) a new object (a red house) after phase 1 and allowed the mouse to explore it for 30 min (Fig. 5a, phase 2). During this second 30 min phase, hyperactivity was still significant in Shank2/−/ mice compared with their wild-type littermates (Wilcoxon rank-sum test; single: \( W = 0, P < 0.001 \); Fig. 5c; paired: \( W = 0, P < 0.001 \); Supplementary Fig. 15a). Again, the level of activity measured in pairs was significantly correlated with the level of activity measured during the single mouse exploration of the test cage (Spearman correlation; phase 2: \( \rho = 0.895, P < 0.001 \)). Interestingly, once the object was introduced, Shank2/−/ mice displayed decreased activity in the object zone (Wilcoxon signed-rank test; single: \( V = 55, P = 0.002 \); Fig. 5b; paired: \( V = 77, P = 0.001 \); Supplementary Fig. 15b). By contrast, Shank2/−/+ mice displayed decreased activity in the object zone (Wilcoxon signed-rank test; single: \( V = 55, P = 0.002 \); Fig. 5b; paired: \( V = 77, P = 0.001 \); Supplementary Fig. 15b), similar to Shank3/−/+ mice (Wilcoxon rank-sum test: \( W = 49, P < 0.001 \); Fig. 5b; and Shank3/−/ mice (\( W = 144, P < 0.001 \); Fig. 5b). This suggests two different exploration strategies: cautious approaches of the object in Shank2/−/+, Shank3/−/+ and Shank3/−/ mice, and uninhibited exploration in Shank2/−/− mice (Fig. 5b, phase 2, bottom row). In paired experiments using the Shank2 strain, the occurrence of each strategy depended on the genotype of the mouse and was not influenced by the genotype of the mouse that they were paired with (see Supplementary Fig. 15a–c).

The two different strategies were characterized by the occurrence of SAP (Fig. 5c,e). SAP is a risk-assessment posture\(^6\) that is automatically quantified in our system, regardless of whether the animals are monitored in single or social conditions. It is characterized by an elongated body (body length longer than the mean body length +1 s.d.) and a reduced speed (<5 cm s\(^{-1}\)). In the object zone, Shank2/−/− mice used the SAP significantly less frequently to explore the novel object zone (Wilcoxon signed-rank test; single: \( V = 72, P = 0.007 \); wild type versus mutant in phase 1; unpaired Wilcoxon rank-sum test; Shank2/−/− versus Shank2/−/+: \( W = 0, P = 4.571 \times 10^{-5} \); Shank3/−/+ versus Shank3/−/−: \( W = 81, P = 2.778 \times 10^{-4} \); wild type versus mutant in phase 2; unpaired Wilcoxon rank-sum test; Shank2/−/− versus Shank2/−/+: \( W = 0, P = 4.571 \times 10^{-5} \); Shank3/−/+ versus Shank3/−/−: \( W = 62, P = 0.100 \) ). Data are mean ± s.e.m. (Shank2/−/+: 10 mice; Shank2/−/−: 8 mice; Shank3/−/+: 7 mice; Shank3/−/−: 12 mice; Wilcoxon rank-sum test; \( ** P < 0.01 \); *** \( P < 0.001 \)). The legend in c also applies to d and f.

Fig. 5 | Hyperactivity and atypical exploration strategy in Shank2/−/− and Shank3/−/− female mice. a. The protocol used when a single mouse was tracked by the system. After 30 min habituation to the test cage (phase 1), a new object (house in red Plexiglas) was introduced for 30 min (phase 2). b. Examples of mouse trajectories during phase 1 and phase 2 for a wild-type mouse, a Shank2/−/− mouse and a Shank3/−/− mouse. Red square, object zone; yellow points, frames in which the mouse is in a SAP. c. Distance travelled in the entire cage during phase 1 and phase 2 for wild-type, Shank2/−/− and Shank3/−/− mice in the single condition (phase 1 versus phase 2; paired Wilcoxon signed-rank test; Shank2/−/−: \( V = 55, P = 0.002 \); Shank2/−/−: \( V = 28, P = 0.195 \); Shank3/−/+: \( V = 28, P = 0.016 \); Shank3/−/−: \( V = 72, P = 0.007 \); wild type versus mutant in phase 1; unpaired Wilcoxon rank-sum test; Shank2/−/− versus Shank2/−/+: \( W = 0, P = 4.571 \times 10^{-5} \); Shank3/−/+ versus Shank3/−/−: \( W = 81, P = 2.778 \times 10^{-4} \); wild type versus mutant in phase 2; unpaired Wilcoxon rank-sum test; Shank2/−/− versus Shank2/−/+: \( W = 0, P = 4.571 \times 10^{-5} \); Shank3/−/+ versus Shank3/−/−: \( W = 62, P = 0.100 \) ). d. Distance travelled in the object zone for wild-type mice and Shank2/−/− mice in phase 1 and phase 2 in the single condition (phase 1 versus phase 2; paired Wilcoxon signed-rank test; Shank2/−/−: \( V = 55, P = 0.002 \); Shank2/−/−: \( V = 0, P = 0.008 \); Shank3/−/+: \( V = 28, P = 0.016 \); Shank3/−/−: \( V = 78, P = 4.883 \times 10^{-5} \) ). e. Definition of the SAP (dotted lines indicate the extended body length of the mouse). f. The proportion of frames in which mice are in SAP in the object zone divided by the total number of frames in which mice are detected in the object zone (wild type versus mutant; unpaired Wilcoxon rank-sum test; Shank2/−/− versus Shank2/−/−: \( W = 78, P = 1.828 \times 10^{-4} \); Shank3/−/+ versus Shank3/−/−: \( W = 77, P = 0.002 \) ). Data are mean ± s.e.m. (Shank2/−/−: 10 mice; Shank2/−/−: 8 mice; Shank3/−/+: 7 mice; Shank3/−/−: 12 mice; Wilcoxon rank-sum test; \( ** P < 0.01 \); *** \( P < 0.001 \)). The legend in c also applies to d and f.
object compared with their Shank2\textsuperscript{−/−} littermates (Wilcoxon rank-sum test; single: \(W = 78, P < 0.001\); paired: \(W = 128, P = 0.001\); Fig. 5f), suggesting that Shank2\textsuperscript{−/−} mice lacked risk assessment and therefore displayed atypical exploration behaviour. The presence of a conspecific did not seem to modulate this abnormality, according to the results obtained in the paired condition. We replicated these findings on hyperactivity and atypical exploration strategy in a second cohort of Shank2 mice that included males and females for single object exploration and females only for paired object exploration (Supplementary Fig. 15d–f). This difference in novel object exploration suggests that Shank2\textsuperscript{−/−} mice present a suppressed neophobia (that is, absence of inhibition); however, this might be independent from their initial increased anxiety in dark–light testing. Notably, Shank3\textsuperscript{−/−} mice also displayed the SAP less frequently in the object zone compared with their wild-type littermates (Wilcoxon rank-sum test: \(W = 77, P = 0.002\)). This suggests that—despite a similar distance travelled around the object—Shank3\textsuperscript{−/−} mice still displayed subtle abnormalities in their exploration strategies.

**Discussion**

LMT makes it possible to phenotype animals in groups and over long periods of time (days to weeks), for both males and females (for an example of a study on males see the Sex-related variations in the behaviour of C57BL/6J adult mice section in the Supplementary Information). The short- and long-term monitoring afforded by the tracking system presented here complement the phenotyping of animals highlighted by classical short-term phenotyping tests (Supplementary Table 3). Indeed, we detected similar variations in activity levels in the short-term object exploration test. In the long-term study, we were able to show that the variations in activity levels in both models were not related to environmental novelty, but were inherent to the mutant mice and persisted through circadian activity. The automatic detection of self-grooming bouts with LMT needs further development, but we were able to document with a high level of details specific social impairments in Shank2\textsuperscript{−/−} mice, as well as the even more subtle social defects shown by Shank3\textsuperscript{−/−} mice. Interestingly, the social deficits displayed by Shank2\textsuperscript{−/−} and even more so by Shank3\textsuperscript{−/−} mice in our long-term experiment were subtle, suggesting that monitoring mice in home cage-like environment might reduce the stress triggered in mice that are tested with one unknown conspecific in a new environment. We also highlighted social deficits that were similar in both strains and were specific to the formation of groups of more than two mice, which is something that could not be evaluated using classical phenotyping methods. Automating behavioural evaluation under any testing conditions (either over long-term or short-term experiments) will improve the robustness of the data collected and eliminate experimenter-related variability in measurements.

This method—which is based on the acquisition of rich sets of data—opens up new avenues for examining how different genotypes, pharmacological tests or enriched environments influence decision-making or social interactions in a robust manner. This system also provides new information on social interactions and, more specifically, on interactions involving more than two freely moving mice. Gathering data on several animals simultaneously and over long periods of time generates large datasets that are highly representative of individual behaviours. This in turn will stimulate massive analysis of large datasets to comprehensively study complex behaviours and should allow a statistical analysis of persistent individual traits. To boost this large-scale approach, LMT provides a website to allow the community to share and analyse scripts, following the example of MouseTube\textsuperscript{21} (see the ‘Collaborative website sharing data’ section in the Supplementary Information; Supplementary Fig. 7).

LMT is a real-time process that also offers new perspectives to interfere with any behaviour online by triggering an external device when a pre-determined event is detected. Indeed, the tracker is able to provide—in real time—the location and the posture information for each mouse. We made this data available by low-latency network connection (see the ‘UDP live network information stream’ section in the Supplementary Information) so that any third-party device (Arduino-like devices enabling automation control) or third-party software can gather current tracking information, either on the computer performing the tracking or on a dedicated computer on a local network. We demonstrate this feature with an example—a live three-dimensional rendering of the subjective view of each mouse created using Unreal Engine (Supplementary Video 5). In this toy demonstration (using only \(x,y,\theta\) for each mouse with \(\theta\) as the orientation angle of the longitudinal axis of the mouse), one can see the scene from the point of view of each mouse. We do not, however, claim an accurate reconstruction of the visual field of the mice. In this demonstration, mice leave a trail of the colour corresponding to their identities to display their past trajectories to better understand the interindividual coordination of movements. This approach can be extended to record electrophysiological activities, ultrasonic vocalizations, physiological signals such as cardiac activity, and also to build closed-loop systems to react to the behaviour of the mice at specific moments with optogenetics or other stimulating systems that are triggered with the closed-loop signals. The system should therefore be used to develop new behavioural tests to better answer the requirements of phenotyping\textsuperscript{22}.

Finally, LMT provides a built-in repertoire of behavioural events that can be analysed using scripts illustrated in this paper. This repertoire includes individual events, such as stop or rearing, but also social interaction events, such as the different types of contact or social configuration involving more than two mice. An exhaustive set of additional data, including the positions \((x,y,z)\) of the head, tail and centre of mass for each animal at each time point, as well as the complete mask of each animal, is available for custom extensive and elaborate analyses. This rich description of individual mice provides the possibility of conducting complex offline analyses on large datasets, such as the analysis of event behavioural sequences\textsuperscript{23}, individual movement tracking or social network analysis. For example, among the Shank2\textsuperscript{−/−} one-day group recordings that we conducted as pilot experiments, we observed a group of four animals that was not nesting in the provided house, whereas all other groups nested there. We investigated this case (Supplementary Video 3). One mouse appeared to avoid the house and refused to enter it. This mouse spent most of its time at the opposite side of the enclosure to the house. The other mice first set up the nest in the house, and then, after a few hours, moved its location and all the nesting material to the location of the other animal. Finally, the structure of the database also allows the addition of extra measurements to monitor the environment for further developments. This should increase the reproducibility of experiments, and be used to optimize housing conditions by analysing behavioural markers of welfare (social isolation and stress) and abnormal behaviour.

The low price of the system enables users to multiplex setups and conduct all experiments at the same time, which is an advantage when running long-term recordings. It is worth noting that the total number of animals that can be tracked simultaneously is only limited by the computing power (each individual tracking has a related CPU cost) and the density of mice present in the field (animals need to be alone from time to time to be identified via RFID). In the future, we plan to connect several setups to provide larger environments for larger groups of mice. We encourage the community to take over LMT and we facilitate ways to improve performances. Indeed, thanks to the open-hardware and software framework—and all the blueprints provided on the web—anyone can reproduce the system, build on it and improve or adapt parts of it without having to painfully rebuild everything from scratch. LMT has been designed natively for parallel architectures and to make the most
out of new and more powerful computing architectures such as the Ryzen CPUs. This scale-up in computing power directly benefits the performances of machine learning experiments, and improves the identification latency. It will also impact the efficacy of data storage and processing during long-term experiments.

Methods

Processing steps of the tracking system. From a processing point of view, the tracking system can be divided into five sections that regroup several sets of processes. For algorithms related to each section, we systematically provide the reference to the code in charge of this process. The first section provides hardware-related information such as camera setup, calibration and RFID reader. The second section describes the processes involved in the detection of the mice in the scene. The third section provides general information about the machine learning process, including motivation of the machine learning algorithm used, features extracted to feed the machine learning algorithm and then its first application within the detection filtering. The fourth section deals with tracking and identity recovery of the tracks. We provide details about the software architecture and how we recover identities of tracks using machine learning, and how we control, correct or set identity results with the RFID readers. Finally, the fifth section describes the head/tail detection features and post-processing.

Machine learning algorithms are used at different times during the procedure. Overall, the machine learning processes use the random-forest algorithm implemented in the weka library (http://www.cs.waikato.ac.nz/ml/weka/) because it complies with the various constraints of our system. There are several machine learning learning threads that are each dedicated to a specific task: one to identity recovery, one to orientation and one to filter out the animals from the background. In our application, the system is not pre-trained because we want to enable adaptation to any type of situation or environment. Also, we require the system to learn rapidly during specific periods of time, and not to keep data from other time segments, to create a specific dataset for which only the mice that are studied are used to train the system. Therefore, we do not use incremental training but rather restart new training using the last set of detections observed in the system. Moreover, we want the learning system to be operational as soon as possible, and to use the smallest number of samples possible to train the dataset. Together, these requirements entail a fast training method and dispense with the need for Support Vector Machine (SVM) deep learning. Random forest is also very well adapted to Adaptive Boosting of data when confidence in detection was low in training sets during the development of the project. The different random-forest threads are used to (1) filter detections, if an object moves, the random forest predicts if the detection is either a mouse or something else (this classification is made with 600 samples per class); (2) predict identity of a detection (5,400 samples per class); and (3) predict their orientation (2,700 samples per class). In all cases, we use 1,000 trees, each with a depth of 100, as parameters. Finally, we do not use boosters in production.

RFID calibration. The reading frequency of the antennas should be at 134 kHz, or 125 kHz, depending on the RFID probes used (the frequency used is dependent on the region). Frequency should be accurate to ±1 kHz to obtain the best reading range. This study and the provided blueprint were designed and performed using 134 kHz probes (glass probe ISO 11784/11785 2×12 mm). The antenna reader enables the self-measurement of the antenna reading frequency in a process called measure unit operating frequency. Use the provided RFIDReaderTest program while soldering to tune its frequency by adjusting wire length. An extra capacitor should be soldered on the reading board to switch to 125 kHz reading.

RFID reader. To read a passive RFID probe carried by an animal, the antenna induces an electromagnetic field in the wave, which powers it. The probe then modulates the received signal to transfer its identification back to the antenna. This whole process is performed in 100 ms by the RFID reader hardware (code at rfid. RFIDAntenna.run). If several probes are powered up at the same time (that is, when several animals are in the range of an antenna), the signal is jammed and no ID is read. The signal is also jammed if several antennas closer than 70 cm are activated at the same time. The RFID identification is in a closed loop with the video tracking.

Segmentation and detection. Capturing the depth map. The depth map is represented as an image of 512 x 424 pixels (typed signed short). We correct invalid values grabbed from the sensor that appear as spikes in the depth map. If the value is within an unexpected range (equals 0 for saturation or less than –32,768), we repeat the value of the previous valid pixel read.

The accuracy of the depth-map measurement is affected by the light absorption of the observed material. The brighter the object is, the further away it artificially appears. The correction is mild (less than 5 mm range), but it is mandatory for observing small animals. We tested eight different Kinects and found a common empirical linear offset correction of the infrared images. We correct the depth map by applying the following formula to each pixel: depthMap = depthMap + (infraValue – 23,000)/1,000.

Computing the background height map. The background height map is represented as an image of 512 x 424 pixels (unsigned short). For each new height map created by the sensor, we update the background height map such that for each pixel, the background height map keeps the minimum value between the captured and the stored maps. Areas containing detections are not processed. Code: livemousetracker.BackgroundHeightMapBuilder.integrateNewDepthImage.

In the case of a false detection event (see below), the height map is corrected using a spurious detection mask. For each pixel, we set the height map values with the values of the detection mask. Thus, the false detection will no longer be detected. Code: BackgroundHeightMapBuilder.correctBackground.

Segmentation and detection. The segmentation map represents all of the animals and objects detected in the field. We compute the Boolean segmentation map as follows—we set the segmentation map as true if heightMap = depthMap > depthSensitivity. In our setup, depthSensitivity was constant and equal to 14, that is, 1.4 mm. If the infrared value is saturated, depth is not reliable and we discard the corresponding pixels from segmentation. We then crop the segmentation mask with the defined region of interest. To obtain all individual segmentations, we perform a connected component extraction (icy internal BooleanMask function). Segmentations are then classified into two lists: spurious and validated detections. Segmentations with less than 30 pixels are assigned to the spurious list (they correspond to sawdust moving in the cage). Segmentations in spurious lists larger than 3 pixels are sent to the backgroundMapBuilder to correct the background. Their count is stored for further analysis in the frame:particle field in the database, as it reflects the sawdust spread by mice when moving, digging or fighting. If the user sets the option ‘Manage wired animal’, extra filtering is applied, based on ellipse fitting of objects. As the longest size of a mouse is around 60 pixels, objects greater than 70 pixels in the third axis are rejected. Then we process the validated segmentation list. If a detection is bigger than MaxSizeDetection, it might be because several animals are in contact, therefore the detection is processed by the detection splitter (see below). Code: detection.MouseDetector.detectMice()
enough information to discriminate between objects. As a constraint, we tried

to find the smallest number of features required to obtain both a good reliability

to the identities of the mice that remain in the solution set. This machine learning

data-mining approach was used only the classes of the mice that are potentially

to be discarded. Roughly, they can be elongated, completely retracted (appearing ball-like),

also an important task. Therefore, we do not want to consider surface, shape descriptor or scale in the learning process. The feature vector is composed of 33 values. The first value is the supervision field—the ID of the animal (that is, its class that is available only when training machine learning, under supervision). The next 16 values are the infrared histogram and the last 16 values the depth histogram.

Code: machinelearning.MachineLearningSetBuilder.

Detection filtering using machine learning. The segmentation process detects all of the objects that are moving and only filters very small objects. However, other objects are moving in the cage as well as the mice, such as food, the house, large areas of dirt, and enrichment objects during object discovery tasks. Therefore, we need an additional filter that is able to determine whether we are observing objects or mice. This filtering is performed by a supervised machine learning algorithm (random forest (1,000 trees, depth 100), weka library).

The machine learning algorithm is trained during the first seconds of the observation, where moving objects are expected to be part of the class Animal versus (such as sponges, boxes labeled as Others. The maximum number of observations considered is 600 per class, which represents the observation of a single mouse for 20s or the observation of four mice for 5s, because they are mixed in this classification. This training set is then updated continuously in a background process that uses the last detection of each animal. Detection filtering is performed for each detection at each frame using a predictor of the machine learning algorithm. If P(detection is class animal) > 0.7 the detection is set as Animal, and otherwise as Others. If the number of items detected remains higher than the maximum number of mice expected, we remove the smallest detections. For each detection, a test is performed to find the number of tracks that terminate close to the detection. If several tracks match a single detection, it means that two or more animals are in contact, so we need to split the detection before the tracking association process starts.

Tracking and identity recovery. The algorithm is based on track reconstruction that consists of detection sequences of identified animals. At the end of the detection procedure, animals are detected but not yet associated with a track nor identified. Different processes are then used to associate a detection with a track and an identity.

Tracking extender association process. Once a detection set is complete, we first try to prolong existing tracks. We obtain all detections at time t - 1. If the detection is less than 30 pixels away, we extend the track. If several tracks are sharing potential prolongations, we perform a Hungarian algorithm to set the best assignment. If no track is found, we create a new track. An important point here is that if the track already has an identity, the associated detection extends this identified track, so the detection takes on a de facto identity.

Code: livemousetracker.Track.TrackExtender.

Software architecture for the tracking identity recovery with machine learning. The tracking identity recovery process is a multi-agent process that is performed as a background task. The tasks addressing the tracks to identify are scheduled by a manager process that uses the last detection of each animal. Detection filtering is performed for each detection at each frame using a predictor of the machine learning algorithm. If P(detection is class animal) > 0.7 the detection is set as Animal, and otherwise as Others. If the number of items detected remains higher than the maximum number of mice expected, we remove the smallest detections. For each detection, a test is performed to find the number of tracks that terminate close to the detection. If several tracks match a single detection, it means that two or more animals are in contact, so we need to split the detection before the tracking association process starts.

Tracking identity recovery with machine learning. An identifier agent receives the track to identify and retrieves concurrent tracks (identified or not) that overlap in time with it. The identifier then attempts to provide the identity of all the tracks involved at once. The first step is to create all of the different possible identity association hypotheses, in a combinatorial manner. This set of solutions is pruned to remove all the hypotheses in which the identities already exist in concurrent tracks. We then create a dedicated machine learning algorithm that is restricted to the identities of the mice that remain in the solution set. This machine learning algorithm is created online, using only the classes of the mice that are potentially involved in the solution, and recent observation data. This provides a more efficient classification algorithm for the location of the mice. This machine learning algorithm is trained live using a random pick of 5,400 detections from each mouse (equivalent to 3 min of observation per animal). We then pick 60 detections from each track, and for each detection we use the machine learning predictor. For each detection of each track, we compute its probability of matching the identity of the hypothesis. Note that we discard the 0 probability, which we floor at 0.01 to avoid a kill switch owing to one potential inferior in the detection set. This provides one probability per detection. We then multiply all of the probabilities together to get the score for one identity association hypothesis. To speed up this process, we cache (that is, store) the custom machine learning created by each hypothesis so that it can be directly queried by another solver without the need to retrain it. Eventually, to adapt any changes in appearance of the mouse or to new configurations displayed, caches are destroyed 2 min after their creation.

In the final step, all of the scores for the different hypotheses are gathered. We retain the best score and compute its share in the sum of all of the scores for all of the detection hypotheses. We take the association decision if the share of the association ratio is greater than 0.95. In that case, we apply the solution found and assign identities to all tracks. If the final association ratio is less than 0.95, the identity is not recovered and anonymous tracks are left as anonymous. The identity manager will therefore start later on a new identifier for those tracks. Meanwhile, the knowledge of the machine learning algorithm with new mice is used to fit the data that remain to be identified, meaning that future observations can solve past tracks. Note that for memory considerations, data are streamed to a database after 5 min, meaning that if no identity is found after 5 min, the track will be saved as anonymous, flushed from memory and will never be identified.

Code: identity.TrackIdentityGlobalSolver.

Triggered RFID closed-loop control for identification. The RFID identification is in a closed loop with the video tracking. The software data structure contains a list of antennas with their respective COM port number, position and range in the image field. It activates priority antennas to query identities of anonymous tracks. We observe all detections available at current t from the anonymous trackpool. We then remove the mice that are in contact in this list as the reader will be jammed if two RFID are present over at the same time. We then activate the antenna that is closest to a detection. Once a reading is performed, we correct the existing tracks and solve possible track conflicts (that is, several tracks that have the same ID at one time point). Machine learning instances operating in the background to solve the identity of the tracks are then cancelled. If no antenna is activated with the previous process (meaning all animals are identified or are all in contact), we use the same procedure to read identities of known animals to constantly validate their identity.

Code: rfid.RFIDManager.activateAntennas.

If an antenna becomes faulty during the process, the user can see it on the interface (the corresponding antenna is coloured red). An antenna can become faulty owing to an unwanted USB unplug or a loss of power. The program communicates with each antenna within its own process to avoid any lock while transmitting data. If a timeout is reached, the antenna is set as faulty. The interface displays the number of reading attempts and the number of effective probe readings for each antenna to monitor the quality of each antenna.

Head/tail post-processing. Masking front and back. Mice mostly move forwards. They can appear to move backwards when they are rearing, during jump phases along walls or when they are running on a wheel. To automatically identify the head/tail position of the mice, we fit an ellipse over the animal and get its major vector. One of the extremities of the ellipse is the location of the head. The mask is then cut into two sub-masks (A and B). We therefore need to identify whether A or B corresponds to the head of the mouse. We use the speed of the animal to set the head mask, we thus need to know in which track this detection is associated to. We calculate its speed. Thus, the head/tail is processed in post-track-processing.

Code: mouseDetection.testA_B_MC_SpeedForHeadLocation().

Swapping the head and tail using machine learning. To perform this recognition, we train one machine learning algorithm per animal (in a background task) using the same feature computation as for the detection. We use 2,780 rolling detections to feed the machine learning algorithm, but we enable its query as soon as we have half of the rolling detection set. As each machine learning algorithm is dedicated to a single animal, they can look completely different, or be equipped with different devices on their head and body. Once the head/tail machine learning of the animal is ready, we compute $PA_A = P(A \text{ is a head})$ and $PBB = P(B \text{ is back of the animal})$ using the predictor $PA_A \times PBB < 1 \text{ and } 1 - PA_A > \sqrt{c_{\text{tt}}} \text{ and } 1 - PBB > \sqrt{c_{\text{tt}}} \text{ with constant (c) } 0.85$, we perform the swap.

Code: mouseDetection.PostProcess().

Mice. P106AP1/Shank2 mutant mice (Shank2−/− mice) were initially described in a previous study. They were bred on a C57BL/6J background (>12 backcrosses),
and were maintained by crossing heterozygous parents. We tested adult female littermates aged between 3 and 12 months of age (the experiments were spread over several months). For the single object exploration, we used 10 Shank2−/− and 8 Shank2+/- female mice aged 4 months of age. In the first cohort, these mice also underwent a previous behavioural characterization that used classical methods (data not shown). The second cohort included 12 Shank2−/− and 7 Shank2+/- female mice, and 8 Shank2−/− and 8 Shank2+/- male mice (aged between 2.5 and 4 months at the time of testing). For the object exploration in pairs, in the first cohort we used 12 Shank2−/− and 12 Shank2+/- mice that were 9 months of age at the time of testing (only two Shank2−/− and four Shank2+/- did not undergo the single object exploration test). We used 13 Shank2−/− and 11 Shank2+/- female mice that were aged between 5 and 6 months from the second cohort (all of these mice underwent the single object exploration task at least three weeks before undertaking object exploration in pairs).

Proband Shank1 mutant mice (Shank3−/− mice) were initially described in a previous study. They were bred on a C57BL/6J background (>10 backcrosses) and maintained by crossing heterozygous parents. We tested adult female littermates (from heterozygous parents). For the single object exploration, we used 7 Shank3−/− and 12 Shank3+/- female mice aged between 2.5 and 3 months of age. We did not conduct the same experiment on the Shank1−/−/− mice.

For the group behaviour experiment in which we used four mice per cage, we constituted social groups at least three weeks before the experiments. We focused our study on female mice (but LMT tracked male and female mice equally well; see the Sex-related variations in the behaviour of C57BL/6J adult mice section in the Supplementary Information). Indeed, constituting mixed-genotype groups of four mice after weaning was not possible for males given their aggressiveness towards unfamiliar same-sex conspecifics at maturity. We constituted nine groups of four mice over the two cohorts of the Shank2 strain (each group included two Shank2−/− and two Shank2+/- mice), which represented 18 Shank2−/− and 18 Shank2+/- mice aged between 3 and 13 months at the time of testing (all of these mice except two Shank2−/− and two Shank2+/- performed the previous experiment). We also constituted six groups of four mice from the Shank3 strain, (each group included two Shank3−/− and two Shank3+/- mice), which represented 12 Shank3−/− and 12 Shank3+/- female mice; these mice were aged between 3 and 4 months at the time of testing but were rested at 10–11 months of age to check for any effects of age. To be able to compare the Shank2 and Shank3 strains, we also tested only adult females in the Shank3 strain. For both strains, the ages were homogenous within each group of four mice for the sake of having age-matched mutant and control mice.

At least three weeks separated two consecutive experiments. Mice were housed in standard laboratory cages in same-sex social groups of two to four mice until the single object exploration. After the single object exploration, for paired object exploration, mice were housed in pairs (wild type and wild type, wild type and mutant, and mutant and mutant; mixing non-littermate mice) at least one week before the experiments. Finally, groups of four mice were constituted at least three weeks before the group monitoring study and were not changed anymore. Overall, social groups were changed a maximum of three times over the course of the experiment pipeline. The Shank2 strain was housed in a 12:12 h dark–light cycle, with lights on at 07:00, whereas the Shank3 strain was housed in an 11:13 h dark–light cycle, with lights turned on at 07:00. Food and water were available ad libitum.

Mice in both strains were identified at weaning (4 weeks of age) using ear punches. The skin sample was used for genotyping by following the protocols described in the original publication. Between two and three months of age, we inserted the RFID tag subcutaneously under isoflurane anaesthetic with local anaesthesia (lidocaine < 0.05 ml at 21.33 mg/ml). All experiments that involved mice complied with the European ethical regulations, and were validated by the ethical committees CETA no 89, Institut Pasteur Paris. Both of these models were characterized behaviourally in previous studies (Supplementary Table 3).

Behavioural protocols. Single and paired object exploration. We placed the experimental cage (50 × 30 × 30 cm) under the setup (70lx, temperature was 22 ± 2°C, noise: fresh bedding (2–3 cm high) covered the bottom of the cage; bedding was renewed for each animal. We placed the tested mouse in the test cage and left it to freely explore the apparatus for 30 min (phase 1). After these 30 min of free exploration, we introduced a novel object (red Plexiglas house, permeable to infrared light; 9.5 × 7.5 × 4.5 cm); Special Diet Services) in the bottom left quarter of the cage. The mouse was left for 30 min in the apparatus (phase 2). In the paired condition of the first cohort, we used three types of pairs (non-littermate pairs): four pairs of Shank2−/− and Shank2+/-, four pairs of Shank2−/− and Shank2+/-, and four pairs of Shank2−/− and Shank2−/−. In the second cohort, we used four pairs of Shank2−/− and Shank2+/-, four pairs of Shank2−/− and Shank2−/−, and four pairs of Shank2+/- and Shank2−/−. These pairs were tested over at least one week before the experiment. We used the software package Python v. 3.6 (http://www.python.org) to compute distances and stretched attend posture in the experiments.

Long-term monitoring of groups of four mice. Nine groups of four mice were created with the Shank2 strain and six groups of four mice were generated with the Shank3 strain. We grouped together two wild-type and two mutant mice at least three weeks before the experiment. We did not control for the fact that wild-type mice may be influenced by the behaviour of the two mutant mice; however, we consider these variations as within the normal variations within a population and therefore we considered the data at the individual level. Each group of four mice was placed in the test cage (50 × 30 × 30 cm; 70lx when the light was on and 0lx when the light was off; temperature was 22°C), with one red house (see above in the single object exploration test), six cylindrical compressed cottons, and food and water ad libitum. Recording started immediately for 24 h for the Shank2 strain and for 3 d for the Shank3 strain. We used Python v. 3.6 to compute distances and behavioural events from the database.

Statistical analyses. We used the software package R 3.3.1 for graphical representation and statistics. For the single and dyadic social interactions, we used non-parametric Wilcoxon rank-sum tests to compare variables between genotypes, given the non-normality of the data and the small sample size. For the long-term recordings of groups of four mice, we used non-parametric Wilcoxon rank-sum tests and Student’s t-tests to compare behavioural events occurrences and duration between genotypes, as well as t-tests to compare the dynamic of the group formation. When necessary, we applied Bonferroni corrections for multiple testing.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The authors declare that all the data supporting the findings of this study are available within the paper and Supplementary Information. Full datasets (databases and films) generated during and/or analysed during the current study are available in the LMT website repository, https://livemousetracker.org.

Code availability
Full source code is available at http://icy.bioimageanalysis.org/plugins/livemousetracker. This includes Java code and also CAD hardware resource files. Python analysis scripts are available at https://github.com/fdechaumont/lmt-analysis.

Received: 11 September 2018; Accepted: 3 April 2019; Published online: 20 May 2019

References
1. Tectot, L. H. & Nestler, E. J. Neurobehavioural assessment in the information age. Nat. Neurosci. 7, 462–466 (2004).
2. Silberman, J. L., Yang, M., Lord, C. & Crawley, J. N. Behavioural phenotyping assays for mouse models of autism. Nat. Rev. Neurosci. 11, 490–502 (2010).
3. Crawley, J. N. Designing mouse behavioral tasks relevant to autistic-like behaviours. Ment. Retard. Dev. Disabil. Res. Rev. 10, 248–258 (2004).
4. Olsson, I. A. S. et al. Understanding behaviour: the relevance of ethological approaches in laboratory animal science. Appl. Anim. Behav. Sci. 81, 245–264 (2008).
5. Belzung, C. & Griebel, G. Measuring normal and pathological anxiety-like behaviour in mice: a review. Behav. Brain Res. 125, 141–149 (2001).
6. Wiltshско, A. B. et al. Mapping sub-second structure in mouse behavior. Neuron 88, 1121–1135 (2015).
7. Schmeisser, M. J. et al. Autistic-like behaviours and hyperactivity in mice lacking ProSAP1/Shank2. Nature 486, 256–260 (2012).
8. By, E. et al. The autism ProSAP1/Shank2 mouse model displays quantitative and structural abnormalities in ultrasonic vocalisations. Behav. Brain Res. 256, 677–689 (2013).
9. Berkel, S. et al. Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation. Nat. Genet. 42, 489–491 (2010).
10. Leblond, C. S. et al. Genetic and functional analyses of SHANK2 mutations suggest a multiple hit model of autism spectrum disorders. PLoS Genet. 8, e1002522 (2012).
11. Leblond, C. S. et al. Meta-analysis of SHANK mutations in autism spectrum disorders: a gradient of severity in cognitive impairments. PLoS One 10, e0140580 (2014).
12. Durand, C. M. et al. Mutations in the gene encoding the synaptic scaffolding gene SHANK3 are associated with autism spectrum disorders. Nat. Genet. 39, 25–27 (2007).
13. Bernardin, K., Elbs, A. & Stiefelhagen, R. Multiple object tracking performance metrics and evaluation in a smart room environment. In Proc. of IEEE International Workshop on Visual Surveillance (2006).
14. Matthews, S. G., Miller, A. L., Plötz, T. & Kyriazakis, I. Automated tracking to measure behavioural changes in pigs for health and welfare monitoring. Sci. Rep. 7, 15782 (2017).
15. Peca, J. et al. Shank3 mutant mice display autistic-like behaviours and striatal dysfunction. Nature 472, 437–442 (2011).
16. Drapeau, E., Riad, M., Kajiwara, Y. & Buxbaum, J. D. Behavioral phenotyping of an improved mouse model of Phelan-McDermid syndrome with a complete deletion of the Shank3 gene. eNeuro 5, e0046-18 (2018).

17. Ohayon, S., Avni, O., Taylor, A. L., Perona, P. & Roian Egnor, S. E. Automated multi-day tracking of marked mice for the analysis of social behaviour. J. Neurosci. Methods 219, 10–19 (2013).

18. Ey, E. et al. Shank2 mutant mice display hyperactivity insensitive to methylphenidate and reduced flexibility in social motivation, but normal social recognition. Front. Mol. Neurosci. 11, 365 (2018).

19. Pappas, A. L. et al. Deficiency of Shank2 causes mania-like behavior that responds to mood stabilizers. JCI Insight 2, 92052 (2017).

20. Holly, K. S., Orndorff, C. O. & Murray, T. A. MATSAP: an automated analysis of stretch-attend posture in rodent behavioral experiments. Sci. Rep. 6, 31286 (2016).

21. Torquet, N., de Chaumont, F., Faure, P., Bourgeron, T. & Ey, E. mouseTube—a database to collaboratively unravel mouse ultrasonic communication. F1000 Res. 5, 2332 (2016).

22. Edser, A. & Grissom, N. Why we need a mouse version of a diagnostic test for autism. Spectrum (21 March 2018); https://spectrumnews.org/opinion/viewpoint/need-mouse-version-diagnostic-test-autism/.

23. Maubourguet, N., Lesne, A., Changeux, J.-P., Maskos, U. & Faure, P. Behavioral sequence analysis reveals a novel role for β2* nicotinic receptors in exploration. PLoS Comput. Biol. 4, e1000229 (2008).

24. R Core Team R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, 2017); http://www.R-project.org/

Acknowledgements
The authors thank Y. Archambeau and P. Ollivon at the workshop of the Institut Pasteur for building the first 12 setups and advising on hardware, W. Meinel for the mathematical proof for decisions of head/tail probability, Microsoft France for their technical support, P. Spinicelli for optical engineering and reading of the paper, R. Marte for machine learning support, B. König for advice and reading of biological experiments, J. N. Crawley for reading and providing comments on the manuscript, A. Barmoutsis for providing us with the early Kinect 2 driver and support, N. Chenouard for driving the use of the machine learning solution, P. Dugast for drawing the mice in the different behavioural events, A. Engelberg for checking the English, S. Wagner and R. Accardi for RFID advice, M. Marim for website development, and X. Montagutelli and M. Bérard for animal facility support. This work was partially funded by the Institut Pasteur, the Bettencourt-Schueller Foundation, the Cognacq–Jay Foundation, the Conny–Maeva Foundation, the ERANET–NEURON SYNPATHY program, the Agence Nationale de la Recherche through grant number ANR-10-LABX-62-IBID, France-BioImaging infrastructure through grant number ANR-10-INBS-04 and the INCEPTION program through grant number ANR-16-CONV-0005, the Centre National de la Recherche Scientifique, the University Paris Diderot, the BioPsy Labex, the Institut National du Cancer through grant number TABAC-16-022, the Foundation for Medical Research (Equipe-DEQ20130326488), the Innovative Medicines Initiative Joint Undertaking through grant agreement number 115300, resources of which are composed of financial contributions from the European Union’s Seventh Framework Program (FP7/2007–2013) and EFPIA companies in kind contribution. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions
F.d.C. created the tracking and analysis methods, developed the open-hardware setup and software. E.E. performed and analysed experiments. N.T. analysed experiments. S.D. provided real-time programming advice. T. Lagache provided the probabilistic framework for machine learning decision-making. T. Legou provided electronic support. A.I. created the blueprints. A.-M.L.S. performed the genotyping of the mice. F.d.C., E.E., N.T., P.F., T.B. and J.-C.O.-M. conceived the project and wrote the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41551-019-0396-1.
Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to F.d., E.E. or J.-C.O.
Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
© The Author(s), under exclusive licence to Springer Nature Limited 2019
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
  - Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  - Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

This work is the initial description of the setup and software used for collecting the data (Live Mouse Tracker, running on the ICY plateform: De Chaumont F, Dallongeville S, Chenouard N, Hervé N, Pop S, Provoost T, et al. Icy: an open bio-image informatics platform for extended reproducible research. Nat Methods. 2012; 9: 690–696. doi: 10.1038/nmeth.2075).

Data analysis

We used Python 3.6 (Python Software Foundation. Python Language Reference, version 3.6; available at http://www.python.org) to compute distances and SAP in the experiments. We used R 3.3.1 for graphical representation and statistics (R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/). Part of the analyses (direction of approaches, group formation) were computed on the ICY plateform (De Chaumont F, Dallongeville S, Chenouard N, Hervé N, Pop S, Provoost T, et al. Icy: an open bio-image informatics platform for extended reproducible research. Nat Methods. 2012; 9: 690–696. doi: 10.1038/nmeth.2075). Analysis codes are available on GitHub.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The authors declare that all the data supporting the findings of this study are available within the paper and its Supplementary Information. Full datasets (databases and films) generated during and/or analysed during the study are available in the Live Mouse Tracker website repository, https://livemousetracker.org. Full source code is available at http://icy.bioimageanalysis.org/plugins/livemousetracker. It includes Java code. This also includes CAD hardware resource files.
Python analysis scripts are available at https://github.com/fdechaumont/lmt-analysis.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

The work was explorative; therefore, we did not estimate the sample size needed. Given the complexity of the behavioral traits examined, we did not have any previous experiments that could provide us with a reasonable effect size to be expected. For the single-object exploration, we studied 10 Shank2+/+ and 8 Shank2-/- female mice in a first cohort; 12 Shank2+/+ and 7 Shank2-/- female mice in a second Shank2 cohort and 8 Shank2+/+; 8 Shank2-/- male mice in a second Shank2 cohort and 7 Shank3+/+ and 12 Shank3-/- female mice. For the paired-object exploration, we used 12 Shank2+/+ and 12 Shank2-/- female mice from the first cohort and 13 Shank2+/+ and 11 Shank2-/- from the second cohort. For the long-term recordings of groups, we tested nine groups of four Shank2 mice, meaning 18 Shank2+/+ and 18 Shank2-/- mice. We also tested six groups of four Shank3 mice, meaning 12 Shank3+/+ and 12 Shank3-/- female mice.

Data exclusions

No data were excluded from the analyses.

Replication

We replicated nine times the experiment of the long-term recordings of a group of four mice for the Shank2 model and six times for the Shank3 model. Data gathered were consistent between all groups within each model.

Randomization

For the single and dyadic object exploration, animals underwent behavioral tests in random order, without knowledge of their genotype. The experimenter was not aware of the genotype at the moment of testing. Genotypes were uncovered only at the end of the analyses. For the long-term recordings, mice were distributed in the different groups according to their genotype (two wild-type and two mutant mice per group). We also avoided (as much as mouse availability allowed) placing littermates in the same group. Genotypes were uncovered only at the end of the analyses computations.

Blinding

Blinding was used in the single and dyadic object exploration tests. Mice were genotyped at weaning, 2 months before the start of the experiments. At the time of experiment, the experimenter was blinded to the genotype of the tested animals. Blinding was not relevant for the long-term monitoring of animals. These animals were recorded without human intervention, only through the tracking system presented in this work.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

- n/a Involved in the study
- Antibodies
- Eukaryotic cell lines
- Palaeontology
- Animals and other organisms
- Human research participants
- Clinical data

Methods

- n/a Involved in the study
- ChIP-seq
- Flow cytometry
- MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

We studied mice mutated in Shank2 or Shank3, two genes associated with ASD risk (Schmeisser et al. 2012 Nature). These mouse lines were generated on a C57BL/6J background. We studied adult females aged between 3 and 12 months of age and adult males aged between 3–6 months. The same animals underwent single-object exploration, dyadic-object exploration and the long-term monitoring study for the Shank2 mice. Shank3 mice underwent single-object exploration and long-term monitoring study (project accepted under the reference 2013-85 by the ethical committee of the Institut Pasteur CETEA n°89).

Wild animals

This study did not involve wild animals.
| Field-collected samples | This study did not involve samples collected from the field. |
|-------------------------|----------------------------------------------------------|
| Ethics oversight        | This project was accepted under the reference 2013-85 by the ethical committee of the Institut Pasteur CETEA n°89. |

Note that full information on the approval of the study protocol must also be provided in the manuscript.