CALL FOR PAPERS | Comparative Genomics

Reliability, robustness, and reproducibility in mouse behavioral phenotyping: a cross-laboratory study


1 Gesellschaft für Strahlenforschung - National Research Center for Environment and Health, Institute of Developmental Genetics, Neuberberg, Germany; 2 Institut Charles Sadron, Strasbourg, France; 3 Medical Research Council, Harwell, United Kingdom; 4 Consiglio Nazionale delle Ricerche - Institute of Cell Biology; and 5 European Molecular Biology Laboratory, Monterotondo, Italy

Submitted 2 April 2008; accepted in final form 23 May 2008

Mandillo S, Tucci V, Höltner SM, Meziane H, Al Banchaabouchi M, Kallnik M, Lad HV, Nolan PM, Ouazzaz A-M, Coghill EL, Gale K, Golini E, Jacquot S, Krezel W, Parker A, Riet F, Schneider I, Marazziti D, Auwerx J, Brown SD, Chambon P, Rosenthal N, Tocchini-Valentini G, Wurst W. Reliability, robustness, and reproducibility in mouse behavioral phenotyping: a cross-laboratory study. Physiol Genomics 34: 243–255, 2008. First published May 27, 2008; doi:10.1152/physiolgenomics.90207.2008.—Establishing standard operating procedures (SOPs) as tools for the analysis of behavioral phenotypes is fundamental to mouse functional genomics. It is essential that the tests designed provide reliable measures of the process under investigation but most importantly that these are reproducible across both time and laboratories. For this reason, we devised and tested a set of SOPs to investigate mouse behavior. Five research centers were involved across France, Germany, Italy, and the UK in this study, as part of the EU-MORPHIA program. All the procedures underwent a cross-validation experimental study to investigate the robustness of the designed protocols. Four inbred reference strains (C57BL/6J, C3HeB/FeJ, BALB/cByJ, 129S2/SvPas), reflecting their use as common background strains in mutagenesis programs, were analyzed to validate these tests. We demonstrate that the operating procedures employed, which includes open field, SHIRPA, grip-strength, rotarod, Y-maze, prepulse inhibition of acoustic startle response, and tail flick tests, generated reproducible results between laboratories for a number of the test output parameters. However, we also identified several uncontrollable variables that constitute confounding factors in behavioral phenotyping. The EU-MORPHIA SOPs described here are an important start-point for the ongoing development of increasingly robust phenotyping platforms and their application in large-scale, multicentre mouse phenotyping programs.

INVESTIGATING MOUSE MODELS of behavior presents challenges in the postgenomic era at many levels: from the initial dissection of the behavioral assay into its composite measures to the identification of phenotypes and their underlying genetic factors, all of which contribute to our understanding of the transient biological processes involved. Accurately attributing the effects of a mutation requires establishing standards for mouse phenotyping, which will be central to the goal of completing the functional annotation of the mouse genome. Over the course of the next few years, a key objective in mammalian functional genomics is to develop a comprehensive library of mutants for every gene in the mouse genome (2, 3). These collections of mutants will ultimately be the focus of systematic phenotyping allowing us to develop a comprehensive view of gene function in the mouse. However, current approaches of screening need to be re-evaluated, particularly with regard to behavioral tests that are highly sensitive to experimental and environmental variables. It will be important to populate databases with phenotype data that are based on common standards and tests and enable robust comparisons to be made between datasets generated from diverse sources (7).

Standardized behavioral screens in mouse genetics are valuable, particularly for large-scale phenotyping enterprises. This conclusion is shared across several laboratories worldwide, and it has been clearly supported very recently by a “Primer” article in Neuron (12). However, setting standards for behavioral phenotyping does not preclude the exploration of novel processes through the development of new platforms and protocols but serves to improve methods in current use for routine effective and accurate screening of mutant lines. There is clearly a need to consider the wider aspects of phenotyping platforms, from collection to subsequent analyses, for meaningful inferences to be made. Nevertheless, efforts aimed at standardizing behavioral phenotyping have brought out both proponents and detractors in the neurosciences. A number of groups (52, 53) have highlighted the potential limitations of standardized batteries of behavioral tests in mice. A notable investigation by Crabbe et al. (10) drew attention to the importance of laboratory environment interaction and other possible confounding factors in replicating behaviors in inbred mice.

The costs of publication of this article were defrayed in part by the payment of page charges. The article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
mouse strains across laboratories. A number of inbred strains and mutants were tested in a battery of six behavioral tests simultaneously across the three laboratories that took part in the study (10). Despite efforts to equate testing equipment, protocols, and animal husbandry methods across laboratories, significant and, in some cases, large effects of site were found (10). However, the scale of the work that will be required to complete a comprehensive functional annotation of the mouse genome underlines the necessity of undertaking phenotyping across multiple centers. Thus reproducibility of results across time and place will be prerequisites for establishing an accurate annotation and represents a challenge that must be overcome if we are to achieve these goals. Standardization is key not only for generating comparable data repositories but also will contribute to reduction and refinement in mouse phenotyping. Employing such standards will enable better characterization of mutant lines and is a platform from which the detailed dissection of experimental manipulations, including environmental and/or testing conditions (19, 43, 47), can be used to assess the effects of the test protocol as well as gene-environment interactions on behavioral outcomes (29, 41).

One of the main objectives of EUMORPHA, a European research project, was the development of a comprehensive genotyping platform that facilitated systematic screening and characterization of mouse mutants with reproducibility (6, 40). The behavioral phenotyping work package (WP10) of EUMORPHA (www.eumorphia.org) focused on the development of a set of standard operating procedures (SOPs) for the detection of relevant behaviors to model neurological and neuropsychiatric disorders and the examination of the SOPs for reproducibility and reliability. We selected a battery of tests that were considered high throughput and informative, covering the following behavioral domains: exploratory drive and anxiety-related behaviors (open field), neurological function (modified SHIRPA, Smithkline Beecham, MRC Harwell, Imperial College, the Royal London Hospital phenotype assessment), muscular strength (grip strength), working memory (spontaneous alternation in a Y-maze), motor coordination and balance (accelerating rotarod), sensorimotor information processing [acoustic startle response (ASR) and prepulse inhibition (PPI)], and pain sensitivity (tail flick). Here we report on results from a stepwise cross-laboratory validation strategy performed on four commonly used inbred strains.

Equipment is commonly considered a possible confound in standardization, yet it is an unrealistic goal to demand such uniform standards in all laboratories. We too regarded this target as unattainable when devising and assessing our standardized battery of tests. A more practical and valuable objective was to employ test apparatus already present in each of our centers to observe the reproducibility of results under conditions that reflect the broad differences in testing regimes that may arise when considering cross-laboratory validation. Furthermore, the aim of standardization in our behavioral battery was not an absolute set of constrictions concerning, for example, housing and equipment, which is neither feasible nor advisable (47, 53), but rather the adaptation of SOPs that were explicitly designed and followed rigorously for the purpose of assessing reproducibility across laboratories.

In summary, the behavioral phenotyping work package in EUMORPHA developed and validated test procedures in four commonly used inbred strains across five research centers for a set of common behavioral tests, employing a minimal range of standards. The study revealed that there are some elements of variability associated with most of the tests, which could account for lack of comparability and poor replication of results previously reported across laboratories in these behavioral platforms. In particular, we found that adopting common standards using well-documented and refined SOPs enabled us to identify and in some cases eliminate common sources of variation, such as test apparatus and experimenter differences, with success. Most notably, the clarification of ambiguous detail within the rotarod test SOP and modifications to the test apparatus resulted in markedly improved reproducibility of results across participating centers.

MATERIALS AND METHODS

Mice and Husbandry

Male mice (n = 10–12) were used only in this study to avoid estrous cycle effects (26). Four research centers [Consiglio Nazionale delle Ricerche (CNR), European Molecular Biology Laboratory (EMBL) Monterotondo, Gesellschaft für Strahlenforschung, and Institut Charles Sadron (ICS)] obtained mice from the same external sources: C57BL/6J, Charles River Germany; C3HeB/FeJ, GSF Germany; BALB/cByJ, Jackson Laboratories US; and 129S2/SvPas, Charles River France. Medical Research Council (MRC) Harwell used in house-derived substrains (C57BL/6J, C3H/HeH, BALB/cAnN, and 129S2/SvEvTac). This allowed us to assess substrain differences within and between centers and to assess the reproducibility of the testing procedures on closely related substrains. BALB/cByJ mice were omitted from the test battery at EMBL Monterotondo, based on quarantine restrictions. Herein, reference to the strains/substrains will be made as follows: C57, C3H, BALB, and 129. In conjunction with standardization, mice were received at 6 wk of age and allowed to habituate for 2 wk, to avoid confounds that transportation may incur on behavior, prior to being put through the test battery, in those centers obtaining mice from external sources. Mice were group-housed (n = 3–5) in a room maintained at a controlled temperature and under a 12:12 h light-dark cycle, with food and water available ad libitum. Testing commenced when the mice were aged ~8 wk. All experiments were carried out in accordance with the guidelines of the US National Institutes of Health, the European Communities Council Directive of 24 November 1986 (86/609/EEC), and local institutional guidelines on the care and use of animals for experimental procedures. Appropriate measures were taken to minimize pain and discomfort of the animals.

Test Battery

A short description of each test equipment and procedures follows; however, detailed SOPs are available on the EUMORPHA webpage: www.eumorphia.org or at http://empress.har.mrc.ac.uk/.

Open field. The open field test measures a combination of locomotor activity, exploratory drive, neophobia, agoraphobia, and other aspects of anxiety and fear in mice. The output of the various interacting drives is locomotion, which is the direct measure obtained. The apparatus consists of a homogeneously and indirectly illuminated angular arena with walls made of plastic material (Table 1).

Mice were placed at the periphery of the open field apparatus with the head facing toward the proximal wall and allowed to explore the arena freely for 30 min. The experimenter was out of view from the mice at all times. The distance traveled and the time spent in the central and peripheral regions were automatically recorded on either a video-tracking system or infrared sensors. The percentage of time spent in the central zone was used as index of emotionality/ anxiety.
**Table 1. Open field test settings in the participating centers**

<table>
<thead>
<tr>
<th>CNR</th>
<th>GSF</th>
<th>ICS</th>
<th>MRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open field dimensions, cm</td>
<td>55 × 32 × 28</td>
<td>48.5 × 48.5 × 38</td>
<td>45 × 45 × 18</td>
</tr>
<tr>
<td>Periphery, cm</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Central zone, %</td>
<td>36</td>
<td>45</td>
<td>40</td>
</tr>
<tr>
<td>Light intensity, Lux</td>
<td>120</td>
<td>180–190</td>
<td>150</td>
</tr>
<tr>
<td>Detector system brand</td>
<td>Video-tracking</td>
<td>Viewpoint</td>
<td>I.R. sensors</td>
</tr>
</tbody>
</table>

CNRS, Consiglio Nazionale delle Ricerche; GSF, ; ICS, Institut Charles Sadron; MRC, Medical Research Council.

Modified SHIRPA. The SHIRPA battery consists of a comprehensive observational phenotypic analysis of the mouse based on a systematic assessment of behavioral and neurological parameters (30, 35). A modified version of the protocol was proposed (6, 40, 41) to reduce the potential for ambiguity and subjectivity within the assessment, as well as to remove duplication where other tests would better evaluate the behavior.

In the modified procedure, each mouse was placed into a viewing jar (5 min) and assessed for unprovoked behaviors, after which mice were transferred to a test arena for a series of observations and manipulations. Score sheets were used to record data semiquantitatively.

The apparatus included: a cylindrical viewing jar (diameter ~15 cm, height ~35 cm) on a raised platform, a Perspex arena (55 × 33 × 18 cm), in which the arena floor was divided into 15 clearly marked squares. A wire grid was secured across the top (middle section) of the arena. A description and list of the scoring parameters as well as a video can be downloaded from the http://empress.har.mrc.ac.uk/ website.

**Grip strength.** Immediately after the modified SHIRPA test, mice were assessed for grip strength performance. Grip strength performance was developed for use in rodent studies (25, 44). It is most commonly used to evaluate the forelimb and hindlimb muscle strength as an indicator of neuromuscular function. Grip strength was measured as tension force using two different types of commercial grip strength meters: Bioseb (Chaville, France), which is designed to measure both forelimb and hindlimb grip strength, and TSE Systems (Bad Homburg, Germany), which measures forelimb grip strength only.

To assess forelimb grip strength measurement, the mouse was held gently by the base of its tail over the top of the grid so that only its front paws were able to grip the grid platform/T-bar. With its torso in a horizontal position the mouse was pulled back steadily until the grip was released down the complete length of the grid/bar. The propensity is that the mouse will cling onto the grid/bar until it can no longer resist the increasing force, before it is released. Grip strength for both sets of limbs was performed similarly (Bioseb only), except with the torso of the mouse parallel to the grid, enabling forelimbs and hindlimbs measurements to be made.

The grip strength meter digitally displays the maximum force applied as the peak tension (in grams) once the grasp is released. The mean of three consecutive trials was taken as an index of forelimb and hindlimb grip strength. Mice were given intertrial interval that varied between centers (see Table 2). Body weight was taken at the end for further analyses.

**Rotarod.** The rotarod is one of the most widely used tests to assess motor coordination and balance in rodents (11, 17). Mice have to maintain their balance on a rotating rod at set or accelerating speeds [e.g., from 4 to 40 revolutions per minute (rpm)]. The latency to fall from the rod is measured for each mouse.

Rotarod was run in two different validation experiments using two different sets of mice. A second validation experiment was necessary because clarification of the initial SOP and slight modification of the equipment were required to enable effective validation. The first experiment was run in conjunction with the test battery; however, the second experiment the test was run in isolation. First validation experiment. Two commercially available rotarod apparatuses were used in three different centers (Table 3) (Letica LEB8200, Panlab, Barcelona, Spain at CNR and ICS, 3375–4 TSE, Homburg, Germany at GSF). Training phase: Mice initially underwent a training session on the apparatus, for three consecutive trials, when the rod was maintained at constant speed. The rod was kept stationary for the first trial and held at 4 rpm for the last two trials. Intertial intervals (ITI) were ~10 min. Providing mice were able to stay on the rod at 4 rpm for 60 s, they were put through the testing phase, with at least 30 min interval between the last training trial and the test phase. The training procedure at GSF differed somewhat; mice underwent training the day prior to the testing phase and two consecutive 180 s trials were performed at 12 and 20 rpm. Test phase: Four trials were completed, with a 15 min ITI. In each trial (T1–T4), four mice were placed on the rod rotating at 4 rpm after which the timer was started and the rod accelerated from 4 to 40 rpm for 300 s. Second validation experiment. The same rotarod apparatus (Letica LEB8200, Bioseb) was used in all three centers participating in the second validation experiment. Clarity in the SOP was required to ensure that each experimenter used the same indexes for scoring. Whilst in both validation experiments, the latency to fall from the rod was determined automatically, the timer was manually stopped if a mouse held onto the rod completing a full rotation (i.e., "passive rotation"). In addition, slight modification was made to the rotarod, at CNR and GSF, in light of the difference in the rod material that was recognized as a variable altering rotarod performance. A soft rubber foam cover was applied to the rod to homogenize this variable. Test phase: The training phase was omitted, and only three trials were completed, with a 15 min ITI. For each trial (T1–T3), only three mice were placed on the rotarod at once, with alternate lanes occupied so that adjacent mice would not influence the performance on the rotating rod. The timer was started when all mice were on the rotarod at 4 rpm, after which it accelerated to 40 rpm for 300 s.

**Table 2. Grip strength test settings in the participating centers**

<table>
<thead>
<tr>
<th></th>
<th>CNR</th>
<th>EMBL</th>
<th>GSF</th>
<th>ICS</th>
<th>MRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equipment brand</td>
<td>Bioseb</td>
<td>Bioseb</td>
<td>TSE</td>
<td>Bioseb</td>
<td>Bioseb</td>
</tr>
<tr>
<td>Sensor module</td>
<td>wire mesh grid</td>
<td>wire mesh grid</td>
<td>T-bar</td>
<td>wire mesh grid</td>
<td>wire mesh grid</td>
</tr>
<tr>
<td>Forelimb test</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Fore- &amp; hindlimb test</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Intertrial interval</td>
<td>10 s</td>
<td>15 s</td>
<td>15 s</td>
<td>7–8 min</td>
<td>5 s</td>
</tr>
</tbody>
</table>

EMBL, European Molecular Biology Laboratory.
Table 3. Rotarod settings used for the first validation in the participating centers

<table>
<thead>
<tr>
<th>CNR</th>
<th>GSF</th>
<th>ICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equipment/brand</td>
<td>Letica LE8200 Bioseb</td>
<td>3375-4 TSE Letica LE8200 Bioseb</td>
</tr>
<tr>
<td>Rod diameter, cm</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Rod material</td>
<td>slightly grooved hard plastic</td>
<td>deeply grooved hard plastic</td>
</tr>
<tr>
<td>Lane width, cm</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 4. Y-maze test settings in the participating centers

<table>
<thead>
<tr>
<th>CNR</th>
<th>ICS</th>
<th>GSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimensions, cm</td>
<td>30 × 9 × 15 opaque black</td>
<td>40 × 9 × 16 clear Plexiglas with specific motifs</td>
</tr>
<tr>
<td>Walls</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oven size, cm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Floor</td>
<td>light grey</td>
<td>light grey</td>
</tr>
<tr>
<td>Light intensity, lux</td>
<td>150–200</td>
<td>100</td>
</tr>
</tbody>
</table>

Y-maze. Spontaneous alternation performance is considered as an index of active retrograde working memory. The Y-maze is used to assess spontaneous alternation and is based on the natural tendency of rodents to explore a novel environment. When placed in the Y-maze, mice generally explore the least recently visited arm and thus tend to alternate their visit between the three arms. Successful alternations included three consecutive visits to a different arm, i.e., the successive arm was not visited immediately prior to the current arm.

Apparatuses consisted of three identical arms placed at 120° from each other, in the shape of a Y, and connected via an equilateral triangular platform in the center. The mouse was placed at the end of one arm, facing away from the center toward the end wall of the arm, and allowed to explore the apparatus freely for 5 min under moderate lighting conditions (100 lux in the center-most region). The initial arm was alternated within the group of mice to prevent bias of arm placement. Latency to leave the first arm and total number and sequence of entries into each arm were scored for each mouse. An arm entry was counted when the mouse had all four paws inside the arm (37). Animals that made fewer than five entries were excluded from the analyses. Y-maze performance was calculated by methods described in current literature (48, 49). A spontaneous alternation was defined as successive entries into each of the three arms on overlapping triplet sets (e.g., ABC, BCA, CAB, etc.). Percentage of spontaneous alternation performance (%SAP) was defined as the ratio of actual alternations (total alternations) to possible alternations (total arm entries − 2) × 100. In addition, total entries were scored as an index of locomotor activity (20) and the latency to exit the starting arm as emotionality-related behavior (data not shown). Table 4 shows Y-maze test settings in the participating centers.

Acoustic startle response and prepulse inhibition. The acoustic startle response (ASR) is characterized by an exaggerated flinching response to an unexpected auditory stimulus. This response can generally be attenuated when it is preceded by a weaker stimulus, the principle underlying PPI. PPI provides an operational measure of sensorimotor gating, which reflects the ability of an animal to integrate sensory information.

ASR and PPI were measured using different types of acoustic startle devices (Table 5). The calibration of the load cell platform amplifier and the white noise tone was performed before each test.

The PPI session was initiated with a 5 min acclimatization period followed by 10 different trial types: one in which the acoustic startle pulse of 110 dB/40 ms was presented only; eight different prepulse trials presented in pseudorandom order in which 10 ms of 70, 80, 85, or 90 dB stimuli either were presented alone or preceded the pulse by 50 ms; and one in which only background noise (BN) was presented that served to measure the baseline movement of the mouse. BN was set at 65 dB. The test session began with five presentations of the acoustic startle pulse alone trial, which were excluded from the statistical analysis. Each acoustic startle, prepulse, or BN trial was then presented 10 times in a randomized order. The ITI was 25 s on average (~20–30 s). The test was conducted with the house lights on. Startle response was recorded every millisecond for 65 ms following the onset of acoustic startle pulse. Maximal peak-to-peak amplitude was used to determine the ASR in the acoustic startle pulse and/or prepulse alone trials. PPI is taken as the prepulse together with startle trial type and expressed as percentage of the basal startle (38).

Tail flick. Pain sensitivity (nociception) is assessed using the tail flick test in rodents (14, 22). The reaction threshold to a high-intensity heat stimulus (acute pain), which is applied to the tail of the rodent, is measured as an index of peripheral pain response. The latency between onset of stimulus and a rapid flick (withdrawal) of the tail from the heat source is automatically recorded.

Two commercially available tail flick apparatus were used (Table 6). Heat intensity was adjusted to produce an individual, stable baseline latency (BL, ~7 s) from which a hypo- or hyperalgesic response can be determined. A threshold was accordingly set, ~3 times the BL response, to prevent tissue damage.

Mice were gently restrained in paper towel cones or using custom-made restraint tubes, in which they were habituated for at least 5 min. The mouse tail was positioned directly under the heat source until it flicked (withdrew). Heat stimulus was applied ~15–25 mm from the tail tip. Three tail flick trials were collected, with an ITI of 1–2 min, from which an average latency was calculated.

Experimental Design

Testing was carried out during the light phase of the light-dark cycle, with at least 1 h between light-dark changes. Figure 1 illustrates a diary that behavioral battery was carried out, the test day and week, and the centers that participated in the test validation. Tests were performed in an order that took into account the nature of behavioral test together with its impact on subsequent tests. For example, the open field test was placed at the start of the behavioral battery since it aims to measure the response to an anxiogenic environment that can be sensitive to prior behavioral test experiences (32). Mice from all four strains were put through the battery of tests in a pseudorandom order and simultaneously to control for possible circadian rhythm effects. At all centers investigators used gloves when handling mice. Experimental apparatuses were wiped clean with water and 50% EtOH solution before each experimental session to prevent olfactory cueing.

Statistical Analysis

Quantitative data were analyzed separately for each test by factorial ANOVA using center and strain as independent variables. Post hoc comparisons were performed when appropriate using Fisher’s LSD test. Qualitative measures (modified SHIRPA) were compared between centers, identifying frequency of different scores in each parameter, statistically by using the χ²-test. The level of significance reported for all comparisons is P < 0.05.

Validation Criteria

Typically, cross-validation is used for studying large data sets and to select prediction models of uncontrolled variables. This procedure
can be used to estimate the significance threshold values to apply to the selection of the “best test” or a combination of acceptable tests from the entire battery. The threshold for acceptance then is decided arbitrarily. The criterion used to consider a test validated in our study is when the order of performance for the strains in that particular test was the same in all three centers.

RESULTS

Open Field

Locomotor and central zone activity within the open field arena revealed differences between strains and centers (Fig. 2). Analysis of variance (ANOVA) indicated that there was a significant strain effect \( P < 0.001 \), center difference \( P < 0.001 \), as well as strain by center interaction (locomotor activity \( P < 0.001 \); central zone activity \( P < 0.01 \)). Interestingly, the strain ranking observed is largely consistent across centers with C57 being the most active strain and 129 least active, while BALB and C3H demonstrate intermediate and comparable levels of activity. The exception is at GSF where BALB were found to be most active. The magnitude of locomotor activity across centers is less stable, and locomotor activity levels as a function of novelty in the four centers distributed as follows: GSF > ICS > CNR > MRC. It is plausible that these differences are a direct index of the differing arena size at the four centers. These effect sizes are also apparent for central zone activity measures, where the length of time spent is proportional to the arena size. Remarkably, however, the order of performance for central zone activity was the same in all centers: C57 > C3H > BALB > 129 (Fig. 2B). Furthermore, comparison of these data shows that in all centers the same overall strain order for performance was observed in at least three strains (C57 > C3H > 129). These results demonstrate that despite differences in equipment at the participating centers, following the clearly devised open field SOP enabled consistent data across inbred strains to be obtained in all the four test centers, whereas the size of effects between centers was relative to equipment differences.

Modified SHIRPA

For the majority of the parameters recorded in the modified SHIRPA test (see www.eumorphia.org for a full description of the parameters) there were few observable differences across laboratories. Differences were found for the following: positional passivity assessment, startle response, touch escape response, trunk curl, and limb grasp scoring (data not shown). A degree of ambiguity in the scoring of these tests was found to be the source of these differences. For example, poor definition of the scale within the modified SHIRPA resulted in differing scores for the startle response at the ICS laboratory. Scoring of the constituent tests was redefined following discussion of the results between centers. It was also observed that subsets of parameters were dependent on the strain; for instance, all parameters recorded above the arena varied between centers for the 129 strain, whereas C3H mice were more consistent across centers. Figure 3 shows the number of square crossings recorded during 30 s of activity in the modified SHIRPA test arena. Strain order across centers was by and large similar. Levels of activity in the SHIRPA test confirmed that the C57 mice were the most active strain in all the centers except at GSF, and the 129 mice were the least active with the exception of MRC. The different 129 substrain used in the MRC may account for the latter difference observed. The modified SHIRPA test was considered to be generally reliable and robust although we have identified some parameters that could be misunderstood or scored inaccurately. Subjectivity within the tests led to the preparation of a SHIRPA video that demonstrates examples of how these specific tests are scored. This video can be downloaded from the EUMORPHIA website and is particularly useful as a tool for training (www.eumorphia.org).

Grip Strength

The scores for grip strength across the five centers are shown in Fig. 4. Note that BALB strain was not tested in one of the centers (EMBL) due to quarantine restrictions. GSF assessed forelimb (2-paws) grip strength only according to their apparatus. Forelimb only analysis of grip strength assessed across centers is found to be relatively consistent with a comparable strain ranking effect: C3H > 129 > C57. A similar pattern is observed with assessment of fore- and hindlimbs combined. These findings are in agreement with a previous study (34) where the C3H strain showed significantly higher grip strength compared with C57 and 129 strains. Some significant differences in the absolute values among centers were apparent when measuring forelimb grip strength. The differences could be the result of animal husbandry methods across centers as the
equipment is reliant on the experimenter applying gentle force. Homogeneity here lies in the technique employed for the test, since it is dependent on the exertion applied, particularly when measuring forelimb grip strength, coupled with the degree at which the mouse is suspended by its tail. Overall the results obtained from the grip strength test demonstrated robust strain differences across all centers. MRC showed these strain effects to a lesser extent that again may reflect strain variation.

**Rotarod**

Development of a reproducible and validated rotarod test across all three participating centers proved to be challenging. Lack of reproducibility and consistency within strains in the first validation effort (Fig. 5A) were possibly due to differences in the apparatus (dimension and material of the rod, see Table 3) used in the three centers. Also, it was recognized that details documented in the SOP were ambiguous and required a coordinated revision to include specific thresholds for latency measures. In the second validation effort all the centers used the same apparatus (LE8200 Letica, Panlab, Spain) with a slight modification of the material used to coat the rotating rod. The SOP was simplified to omit the initial training phase and reduce the trials from four to three; a “passive rotation” was defined as an end point leading to termination of the experiment; and the use of a foam cover on the rod facilitated performance consistency. Remarkably, results from the second validation demonstrate a marked improvement in both the reproducibility and strain ranking effects across all centers. C57 and BALB appeared to be most sensitive to test center differences, whereas C3H and 129 were less affected.

**Y-maze**

Percentage of spontaneous alternation performance (%SAP) in the Y-maze revealed no significant differences across centers for either strain or strain × center interaction (Fig. 6A). Interestingly, few clear strain effects were apparent in %SAP, and only the 129 strain performed comparably in all centers, demonstrating the least %SAP. In contrast, significant differences were observed in arm entries for strain by center interactions (Fig. 6B). However, a robust effect was again seen in 129 strain that were consistently making fewer arm entries across test centers. These findings corroborate our observations for locomotor activity in this strain within the open field and modified SHIRPA tests, which were reduced compared with the other strains. Absolute values for arm entries appeared to be marginally lower at ICS that may have been a direct result of differing dimensions, arm length of Y-maze is 10 cm greater than the other two centers. The overall lack of strain ranking, across the C57, BALB, and C3H strains and the three centers, is less easy to comprehend because the test is not highly reliant on animal husbandry methods or obviously sensitive to equipment differences, and analysis is routine. Differences here may be more intrinsic to the parameters of the assay, and the vagaries of its reliability to inform us of active retrograde working memory.

**ASR and PPI**

The startle reflex varied in magnitude and across strains between centers (Fig. 7A), particularly regarding the varying acoustic stimulus intensities. Most of these differences are most likely to be explained by the alternative acoustic devices used at the centers. ICS and CNR employed the same branded device (SR Lab) and obtained a similar magnitude of results, whereas the other two centers used very disparate devices that generated incongruous results. While the settings for experimentation were standardized across centers where possible, there are a number of parameters that could not be equated, such as the soundproof chamber, sensor sensitivity, as well as the inherent characteristics of white noise, which may have introduced the qualitative and quantitative variability observed. For example, the prepulse stimulus intensity at GSF required an equivalently higher setting due to a difference in the baseline BN, i.e., the weakest prepulse corresponded to 30 dB rather than 5 dB above the BN. This could also explain the maximal level of PPI reached at GSF with the weaker prepulse of 70 dB (Fig. 7B). Despite the acoustic startle differences, interstrain comparisons for PPI were highly robust across the four centers, especially at a prepulse of 80 dB, which produced a consistent magnitude of response (Fig. 7B). The two centers that used the same apparatus (CNR and ICS) obtained almost identical numerical values in three out of four strains for global PPI collapsed across all prepulse intensities from 70 to 90 dB: BALB, 50 ± 4 and 29 ± 6; C57, 44 ± 3 and 39 ± 4; 129, 78 ± 2 and 75 ± 1; C3H, 59 ± 4 and 58 ± 3 for ICS and CNR, respectively. Remarkably, the strain order we found in our test
battery (129 > C3H > C57 and BALB) is consistent with that reported in the literature (31, 33). Overall it is clear that fewer disparities are observed in PPI across laboratories and time than in the ASR.

**Tail Flick**

Figure 8 illustrates the mean latency for tail flick in the four strains across the three participating centers. Note that BALB strain was not tested in one of the centers (EMBL) due to quarantine restrictions. ANOVA and post hoc analysis detected significant differences between centers. Nevertheless, strain ranking effects were generally similar (C3H > C57 > 129) with variability of the 129 mice only observed in one center. Similar strain differences (i.e., C57BL/6J more sensitive than C3H/HeJ) have also been reported in a study in which tail withdrawal was used to evaluate nociception in different mouse strains using various pain assays (27). In addition, when only the data from CNR and ICS were compared, the ANOVA showed that the test center-by-strain interaction was no longer significant (F2,64 = 1.10; P = 0.3388). The potential confound with the tail flick paradigm as it stands is excessive handling required to restrain the mouse onto the heat source that may be an additive stressor. Mouse performance may be subsequently...
The ANOVA revealed a statistically significant effect of strain [F(2, 91) = 77.60, P < 0.0001], center [F(3, 117) = 14.79, P < 0.0001], and a statistically significant strain × center interaction [F(6, 117) = 3.61, P = 0.0025]. Statistically significant differences between centers within each strain: *P < 0.05 CNR vs. EMBL, GSF vs. CNR; **P < 0.005 GSF vs. ICS, CNR vs. GSF; ***P < 0.0001 ICS vs. CNR, EMBL vs. CNR. Statistically significant differences between strains within each center: +P < 0.05 C57 vs. 129, C3H vs. 129, ###P < 0.0001 C57 vs. C3H, 129 vs. C3H, Fisher’s PLSD test. 

In conclusion, we have collated and summarized in Table 7 the key results reflecting the main effects and interactions from the ANOVA performed on each test as well as the strain ranking consistencies across centers. This analysis shows that many tests show very significant strain effects. Moreover, a number of the tests show very significant center effects. Importantly, however, a large number of tests show consistent strain ranking across centers.

FIGURE 4

Grip strength test

Forelimb (2 paws) and fore-/hindlimb (4 paws) grip strength measurements of 3 mouse strains in 5 Eumorphia centers (CNR, EMBL, GSF, ICS, and MRC). Bars represent the means (±SE) grip strength measure-

FIGURE 5

Rotarod test. Rotarod performance of four inbred strains before (A) and after (B) optimization of apparatus and procedure in 3 EUMORPHIA centers (CNR, GSF, and ICS). Bars represent means (±SE) latency to fall from rotating rod (4–40 rpm in 300 s) averaged across all test trial. A: first validation experiment. The test (4 trials) was preceded by a training phase, and 3 different apparatus were used in the centers. The ANOVA on the 1st validation data revealed a statistically significant effect of the center [F(2, 121) = 86.64; P < 0.0001], strain [F(3, 121) = 6.69; P = 0.0003], and as well as of the interaction center × strain [F(6, 121) = 5.37; P < 0.0001]. Fisher’s post hoc test, LSD between centers within each strain: *P < 0.05 vs. GSF, **P < 0.005 vs. CNR, ***P < 0.0001 CNR, ICS; and between strains within each center: +P < 0.05 vs. C57, BALB, C3H, 129. Fisher’s post hoc LSD did not measure 4-paw grip strength.

In conclusion, we have collated and summarized in Table 7 the key results reflecting the main effects and interactions from the ANOVA performed on each test as well as the strain ranking consistencies across centers. This analysis shows that many tests show very significant strain effects. Moreover, a number of the tests show very significant center effects. Importantly, however, a large number of tests show consistent strain ranking across centers.

DISCUSSION

One of the major challenges that we are faced with in successfully annotating the mouse genome and fully characterizing mouse mutants is the development of reliable high-throughput phenotyping platforms that allow the acquisition of comparable datasets, which will in turn effectively populate
unified and integrated databases. Supported by the European Commission, under Framework 5 and 6 [EUMORPHIA (QLG2-CT-2002-00930) and EUMODIC (LSHG-CT-2006-037188)], we devised a standardized first-line behavioral phenotyping screen to assess and validate a set of SOPs for reproducibility across laboratories and time. Despite the use of a range of equipment across the participating centers, our study demonstrates that it is possible to achieve strikingly reproducible and robust strain effects, as well as revealing several confounds that limit absolute replication in specific tests. In particular, the sources of variability associated with individual tests were mostly found to be the result of: (1) experimenter experience and animal husbandry methods (SHIRPA, grip strength, tail flick test); (2) apparatus differences (open field, acoustic startle, rotarod, Y-maze test); and (3) clarity within the SOP (SHIRPA and rotarod).

Strain ranking effects from the open field data generated in our test battery are in agreement with previous reports (10). Several laboratories have replicated the finding that C57BL/6J is more active compared with 129/Sv in different test situations, including spontaneous locomotor activity, light-dark exploration, and open field (5, 45). Carola et al. (8) have also reported that C57BL/6J displayed greater activity than BALB/c and C3H/He strains in open field, plus maze and free exploratory tests. Interestingly, these findings extended to the use of substrains in our study indicating robust phenotypic differences. While some studies have found that C57BL/6 is less anxious than BALB/c (1, 13) and 129 strains (45), data from other tests classically used for the evaluation of anxiogenic behaviors, such as the light-dark test and elevated plus maze, or trait anxiety, such as the free exploratory test, globally suggest a strain order C57BL/6 < C3H < BALB/c from least to most anxious (4, 16). We showed here that despite differences in equipment and discrete environmental conditions, the SOP devised for the open field test provided reliable and consistent results between the four test centers, with a homogenous strain order of performance for either locomotor activity and/or anxiety-related parameters. The magnitude of absolute values for each of these parameters did differ between centers; however, these appear to equate to differences in equipment, whereby arena size was proportional to the amplitude of locomotor activity and percentage center time. The lower levels of locomotor activity, as a function of novelty, observed at the MRC and CNR could be explained by the smaller arena used for the open field test that also resulted in an increased time spent in the center.

Evaluation of experiences across centers allowed us to conclude that inadequate detail within the modified SHIRPA SOP as well as the categorical nature of scoring, which resulted in variable phenotype assessments, confounded reproducibility across test centers for a few of the scores. To improve this experimenter variability, a demonstration movie file was made that explicitly differentiated between some specific categorical scores where ambiguity was apparent. Experimenter variability was also a factor in the grip strength test. Relatively robust strain effects were observed in this test; however, absolute values were reliant on the technique employed such that experimental force was a factor when forelimb strength was being measured.

Rotarod was particularly useful in highlighting the potential for refinement of SOPs during our validation exercise, after which we were able to produce comparable results across the participating institutes. A careful redefinition of the SOP and modification of the material covering the rotating rod was necessary to attain a level of comparability across centers. Interestingly, this test has been used widely to evaluate motor deficits in models of human pathological conditions such as Parkinson’s and Huntington’s diseases. Despite its popularity, this test has been contentious in obtaining reproducible results and often yielded discordant data across time and laboratories (10). The lack of reliability for this test is likely due to varying
attributes of the test procedure used and qualitative differences in equipment (36). Our results largely support this notion. Rank order of performance showed a good degree of reproducibility and robustness for the strains included in our battery and replicated findings in other studies (23, 38).

The Y-maze test showed poor consistency across strains and centers. Absolute values differed between centers possibly as a result of the dimensional facets of the equipment. The robust strain effect seen in the 129 strain substantiated the lower locomotor activity found for this strain in the open field and modified SHIRPA tests. Nonetheless, for the other strains there was poor reproducibility of strain differences across centers. Given that the outcome of this assay is less dependent on husbandry effects, it is difficult to effectively dissect the potential source of variation. It is reasonable to conclude that the inherent indexes of the assay that are used to inform us of active retrograde working memory are questionable. In fact, some concerns have been raised about the reliability of this test with reference to assessing working memory (15). Our results might suggest that mice of the 129S2/SvPas strain demonstrate impaired spatial navigation in the Y-maze compared with the other three strains, in line with a previous finding that 129 mice are poor learners compared with the C57BL/6 strain (51). However, there is evidence that some 129 mouse substrains do not differ in spatial navigation compared with C57BL/6 mice (28). The low-active characteristic of the 129 strain plausibly contributes to the poor response in the Y-maze test, and so caution is warranted in making generalizations about strain differences with respect to this test procedure (42).

Fig. 7. Acoustic startle response and prepulse inhibition (PPI). The figure illustrates startle responses (A) and PPI (B) data obtained in the 4 testing centers, CNR, GSF, ICS, and MRC. Data are expressed as means ± SE. A: startle reactivity to the prepulses and the 110-dB pulse. BN, background noise (65-dB). Three-way ANOVA for startle responses revealed a statistically significant main effect of center [F(3,155) = 224.34, P < 0.0001], Strain [F(3,155) = 24.60, P < 0.0001], Sound intensity [F(3,465) = 1483.80, P < 0.0001], and a statistically significant center × strain × sound intensity interaction [F(45,775) = 6.00, P < 0.0001]. The absolute numerical values of the startle response to the main pulse (110 dB) varied between centers, but the overall pattern of the strain differences tended to be comparable. Fisher’s post hoc test, LSD between strains within each center: *P < 0.05 vs. BALB; ++P < 0.01 vs. C57, BALB; +++P < 0.005 vs. 129; ++++P < 0.0001 vs. all. B: percent PPI obtained with the 2 weak prepulses, 70 and 80 dB. The numerical values of %PPI varied for each mouse strain depending on the prepulse intensity and testing centers as indicated by the statistically significant strain × prepulse level × center interaction [F(27,465) = 3.56, P < 0.0001]. Nevertheless, the overall pattern of the strain distribution tended to be consistent across the centers, especially for the 80 dB prepulse. Indeed, the strains order for the 80 dB prepulse was remarkably similar in CNR, GSF, and ICS, with both C57BL and BALB mice having the lowest scores and 129 mice the greatest scores. Fisher’s post hoc test, LSD between strains within each center: *P < 0.05 vs. C57; ++P < 0.05 vs. BALB; ++P < 0.005 vs. C3H; +++P < 0.005 vs. 129; ++++P < 0.0001 vs. C3H; +++++P < 0.0001 vs. C57 and/or BALB; ++++++P < 0.0001 vs. all.
The ASR and PPI paradigms are widely used in a number of clinical studies since it has been shown that PPI deficits are present in several psychiatric conditions, most notably schizophrenia. The lack of sufficient sensory gating control is thought to lead to an overflow of sensory stimulation and disintegration of cognitive functions in these patients. One of our aims in devising the ASR and PPI SOP was to determine to what extent PPI is amenable to standardization. In mice it is often used to assess psychiatric and neurodegenerative disease models (such as schizophrenia and Huntington’s disease). It is critical therefore that accurate deviation from baseline levels can be measured to make any valid inferences. The qualitative and quantitative differences in equipment potentially determined the amplitude of the ASR in our study. Two of the four participating centers, employing the same measuring device, produced similar absolute results. In contrast, the other two centers’ results were less congruous. This raises some concern regarding the wider implications of the startle response test when used to assess the effects of putative antipsychotics and to explore genetic and neurobiological mechanisms underlying behaviors relevant to psychosis (31). The magnitude of the response may be dependent on the device used rather than the drug dose and/or mutation effect. PPI on the other hand was less influenced by these factors and produced remarkably reproducible and significant strain differences across centers.

The need for standardization of a test to assess pain sensitivity in genetically engineered mice has been discussed at length (50). In our hands, the tail flick test generated relatively robust strain effects despite the test being largely reliant on animal handling methods that involve the experimenter exerting initial restraint on the mouse under test. While time is allowed for the mouse to habituate to the restraining device, some strains of mice (e.g., BALB/c) are more sensitive to handling and restraint that may influence the BL for tail withdrawal (9). Overall, however, we found strain differences to be reasonably reproducible with the exception of the 129 in one test centre.

Baseline strain differences have been widely reported for behavioral tests (11). However, what we have demonstrated in our study is that it is possible to measure robust strain differences across laboratories, which are to some extent test specific. In particular, the C57 and 129 strains consistently showed a high level of comparability across centers and often were found to be at the extremes on a performance scale in almost all the tests: C57 high-active and 129 low-active. Conversely, the performance of BALB was least comparable across centers, which could be an intrinsic characteristic of this strain. These results lead to an unpredictable strain ranking order throughout the test battery for the intermediate strains, BALB and C3H.

The use of a test battery serves to obtain high-throughput data, reducing the number of animals required and providing a multidimensional evaluation of individual mouse lines. Nevertheless, there are potential problems involved with multiple testing, and test order can influence behavioral outcomes (24, 32). We accept that the findings reported here could differ if the same SOP was employed in isolation and not as part of a test battery. However, we believe that this is unlikely given the carefully devised test order. The test sequence and the intervals between tests in our battery were selected with the aim of providing a summary of the consistency of strain rankings across centers. SHIRPA, Smithkline Beecham-MRC Harwell-Imperial College-Royal London Hospital phenotype assessment; %SAP, percentage of spontaneous alternation performance.

The table illustrates the level of significance of main effects (strain or center) and 2-way interaction (strain × center) from factorial ANOVA as well as providing a summary of the consistency of strain rankings across centers. SHIRPA, Smithkline Beecham-MRC Harwell-Imperial College-Royal London Hospital phenotype assessment; %SAP, percentage of spontaneous alternation performance.

<table>
<thead>
<tr>
<th>Test</th>
<th>Measure</th>
<th>Strain</th>
<th>Center</th>
<th>Strain × Center</th>
<th>Strain Ranking Consistency (# of Centers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open field</td>
<td>distance traveled</td>
<td>P &lt; 0.0001</td>
<td>P &lt; 0.0001</td>
<td>P = 0.0012</td>
<td>C57 &gt; C3H, BALB &gt; 129 (4/4)</td>
</tr>
<tr>
<td>Open field</td>
<td>% time in center</td>
<td>P &lt; 0.0001</td>
<td>P &lt; 0.0001</td>
<td>P &lt; 0.0001</td>
<td>C57 &gt; C3H &gt; BALB &gt; 129 (4/4)</td>
</tr>
<tr>
<td>SHIRPA</td>
<td>crossings</td>
<td>P &lt; 0.0001</td>
<td>P &lt; 0.0001</td>
<td>P &lt; 0.0001</td>
<td>C57 &gt; C3H, BALB &gt; 129 (4/5)</td>
</tr>
<tr>
<td>Grip strength</td>
<td>2-paw force</td>
<td>P &lt; 0.0001</td>
<td>P &lt; 0.0001</td>
<td>P = 0.0025</td>
<td>C57 &lt; 129 &lt; C3H (5/5)</td>
</tr>
<tr>
<td>Grip strength</td>
<td>4-paw force</td>
<td>P &lt; 0.0001</td>
<td>P = 0.039</td>
<td>P = 0.229</td>
<td>C57 &lt; 129 &lt; C3H (3/4)</td>
</tr>
<tr>
<td>Rotarod 1&lt;sup&gt;st&lt;/sup&gt; validation</td>
<td>latency to fall</td>
<td>P = 0.0001</td>
<td>P &lt; 0.0001</td>
<td>P &lt; 0.0001</td>
<td>not consistent</td>
</tr>
<tr>
<td>Rotarod 2&lt;sup&gt;nd&lt;/sup&gt; validation</td>
<td>latency to fall</td>
<td>P = 0.0001</td>
<td>P &lt; 0.0001</td>
<td>P &lt; 0.0143</td>
<td>C57 &gt; 129 &gt; C3H (3/3)</td>
</tr>
<tr>
<td>Y-maze</td>
<td>% SAP</td>
<td>P = 0.0006</td>
<td>P &lt; 0.0001</td>
<td>P &lt; 0.0001</td>
<td>129 &lt; C57, C3H, BALB (3/3)</td>
</tr>
<tr>
<td>Y-maze</td>
<td>tot entries</td>
<td>P = 0.0001</td>
<td>P &lt; 0.0001</td>
<td>P &lt; 0.0001</td>
<td>129 &lt; C57, C3H, BALB (3/3)</td>
</tr>
<tr>
<td>ASR-PPI</td>
<td>startle response</td>
<td>P &lt; 0.0001</td>
<td>P &lt; 0.0001</td>
<td>P &lt; 0.0001</td>
<td>C57 &lt; C3H, BALB, 129 (4/4)</td>
</tr>
<tr>
<td>ASR-PPI</td>
<td>% PPI</td>
<td>P &lt; 0.0001</td>
<td>P &lt; 0.0001</td>
<td>P = 0.0037</td>
<td>C57, BALB &lt; C3H &lt; 129 (4/4)</td>
</tr>
<tr>
<td>Tail flick</td>
<td>latency</td>
<td>P = 0.0014</td>
<td>P = 0.0001</td>
<td>P = 0.0006</td>
<td>C3H &gt; C57 &gt; 129 (2/3)</td>
</tr>
</tbody>
</table>

Fig. 8. Tail flick test. Mean latency to flick tail from heat source averaged across test trials in 3 inbred mouse strains tested in 3 EUMORPHIA centers (CNR, EMBL, ICS). Bars represent means ± SE. Data from the BALBcByJ strain were not included in the analysis because they were not tested in 1 center due to quarantine limitations. In the other 2 centers however, the mean latencies (±SE) from these mice were significantly different (CNR: 7.76 (±0.52) s; ICS: 14.96 (±0.47) s, Fisher’s post hoc test, P < 0.0001). The ANOVA for the 3 remaining strains compared across the 3 testing centers revealed statistically significant effects for the main factors center [F(2,90) = 40.07, P < 0.0001], strain [F(2,90) = 7.08, P = 0.0014], as well as of the 2-way interaction center × strain [F(4,90) = 5.40, P = 0.0006]. Statistically significant differences between test centers within each strain: *P < 0.01 CNR vs. ICS, EMBL vs. ICS; **P < 0.001 EMBL vs. ICS, CNR vs. ICS; ***P < 0.0001 EMBL vs. CNR, ICS. Statistically significant differences between strains within each center: #P < 0.05 C57 vs. 129, ###P < 0.005 C3H vs. 129, ####P < 0.0001 C3H vs. 129, +P < 0.01 C57 vs. C3H, Fisher’s PLSD test.
reducing confounding influences between tests as much as possible. McIlwain et al. (24) have recently shown that prior testing experience, handling, and exposure to a behavioral test battery had little or no effects on anxiety-related behaviors compared with naive mice. In fact, it is plausible that handling in preceding tests may invoke a reduced habituation response (46).

In summary, we have developed SOPs for neurological and behavioral studies and studied the reproducibility of a behavioral test battery. Our findings demonstrate that it is possible to standardize tests using well-defined SOPs and wide-ranging equipment, and we were able to uncover robust strain effects. Key factors that could be responsible for baseline differences reported for some of these behavioral tests have been highlighted and suggested. As we seek continued improvements in SOP reproducibility, it will be important to consider the variables identified here in the ongoing design process. Specifically, we demonstrate when comparing tests performed in different laboratories that absolute values are less reproducible while strain ranking effects are a better index of replication (see Table 7). Nevertheless, this augurs well for future large-scale studies at multiple centers. Analysis of the phenotype of mutants will depend on identifying significant differences to the baseline control levels as well as assessing and identifying outliers from the normal range, equivalent to determining strain ranking. Each center will determine wild-type control baseline levels, and if tests are used that reproducibly identify outliers from the normal range then we can expect to develop valid and comparable datasets summarizing mutant phenotypes that are not confounded by center.

With the advent of novel approaches in phenotyping methods, it is clear that use of automated equipment that remove subjectivity of scoring and excessive experimenter handling will aid in our efforts to gain homogeneous data sets; however, these are ultimately dependent on standardization of criteria set. Automated home cage systems have largely been successful in studying behavioral and cognitive processes in mice (18, 21, 39). These advances have the potential for defining specific experimental variables with relative accuracy, thus further reducing the source of confounding factors and generating high-throughput phenotyping data.

In conclusion, the SOPs developed in the EUMORPHIA program provide an important start point for comprehensive phenotypic analysis of mouse mutants across centers. Data collected in this study are already serving as a foundation for launching another EU-funded project, EUMODIC (http://www.eumodic.eu/), which is undertaking the comprehensive primary phenotype assessment of up to 650 mouse mutant lines.

REFERENCES


