Using mice from different breeding sites fails to improve replicability of results from single-laboratory studies

Jaric, Ivana ; Voelkl, Bernhard ; Amrein, Irmgard ; Wolfer, David P ; Novak, Janja ; Detotto, Carlotta ; Weber-Stadlbauer, Ulrike ; Meyer, Urs ; Manuella, Francesca ; Mansuy, Isabelle M ; Würbel, Hanno

DOI: https://doi.org/10.1038/s41684-023-01307-w

https://doi.org/10.1038/s41684-023-01307-w

Using mice from different breeding sites fails to improve replicability of results from single-laboratory studies

Ivana Jaric[1] ✉, Bernhard Voelkl[1], Irmgard Amrein[2], David P. Wolfer[2,3], Janja Novak[1], Carlotta Detotto[4], Ulrike Weber-Stadlbauer[5], Urs Meyer[5], Francesca Manuella[6,7,8], Isabelle M. Mansuy[6,7,8] & Hanno Würbel[1] ✉

Theoretical and empirical evidence indicates that low external validity due to rigorous standardization of study populations is a cause of poor replicability in animal research. Here we report a multi-laboratory study aimed at investigating whether heterogenization of study populations by using animals from different breeding sites increases the replicability of results from single-laboratory studies. We used male C57BL/6J mice from six different breeding sites to test a standardized against a heterogenized (HET) study design in six independent replicate test laboratories. For the standardized design, each laboratory ordered mice from a single breeding site (each laboratory from a different one), while for the HET design, each laboratory ordered proportionate numbers of mice from the five remaining breeding sites. To test our hypothesis, we assessed 14 outcome variables, including body weight, behavioral measures obtained from a single session on an elevated plus maze, and clinical blood parameters. Both breeding site and test laboratory affected variation in outcome variables, but the effect of test laboratory was more pronounced for most outcome variables. Moreover, heterogenization of study populations by breeding site (HET) did not reduce variation in outcome variables between test laboratories, which was most likely due to the fact that breeding site had only little effect on variation in outcome variables, thereby limiting the scope for HET to reduce between-lab variation. We conclude that heterogenization of study populations by breeding site has limited capacity for improving the replicability of results from single-laboratory animal studies.

Experimental animal research is usually conducted using animals of the same genotype (inbred or mutant strains) reared and housed under almost identical conditions[1]. Such rigorous genetic and environmental standardization can produce study-specific results that lack external validity[2–4], thereby causing poor replicability[5–7]. Theoretical and empirical evidence indicates that systematic heterogenization of study populations, rather than standardization, is needed to improve external validity and replicability[6,8–11]. However, previous studies indicate that simple forms of heterogenization (for example, varying cage size, group size, environmental enrichment or including multiple experimenters) are not effective enough to attenuate the large heterogeneity that normally exists between independent replicate studies[8,12,13]. Therefore, there is a need for more effective ways of heterogenizing study cohorts within single-laboratory studies to generate results that are replicable across independent laboratories.

We recently found that common environmental differences between animal facilities produce facility-specific phenotypes in mice, from the molecular to the behavioral level[14]. These findings suggest that the animals'

[1]Animal Welfare Division, Vetsuisse Faculty, University of Bern, Bern, Switzerland. [2]Institute of Anatomy, Division of Functional Neuroanatomy, University of Zürich, Zürich, Switzerland. [3]Department of Health Sciences and Technology, ETH Zürich, Zürich, Switzerland. [4]Central Animal Facilities, Experimental Animal Center, University of Bern, Bern, Switzerland. [5]Institute of Pharmacology and Toxicology, Vetsuisse Faculty and Center of Neuroscience Zürich, University of Zürich, Zürich, Switzerland. [6]Laboratory of Neuroepigenetics, Brain Research Institute, Medical Faculty, University of Zürich, Zürich, Switzerland. [7]Institute for Neuroscience, Department of Health Science and Technology, Swiss Federal Institute of Technology Zürich (ETHZ), Zurich, Switzerland. [8]Center for Neuroscience Zürich, University Zürich and ETHZ, Zürich, Switzerland. ✉e-mail: ivana.jaric@unibe.ch; hanno.wuerbel@unibe.ch

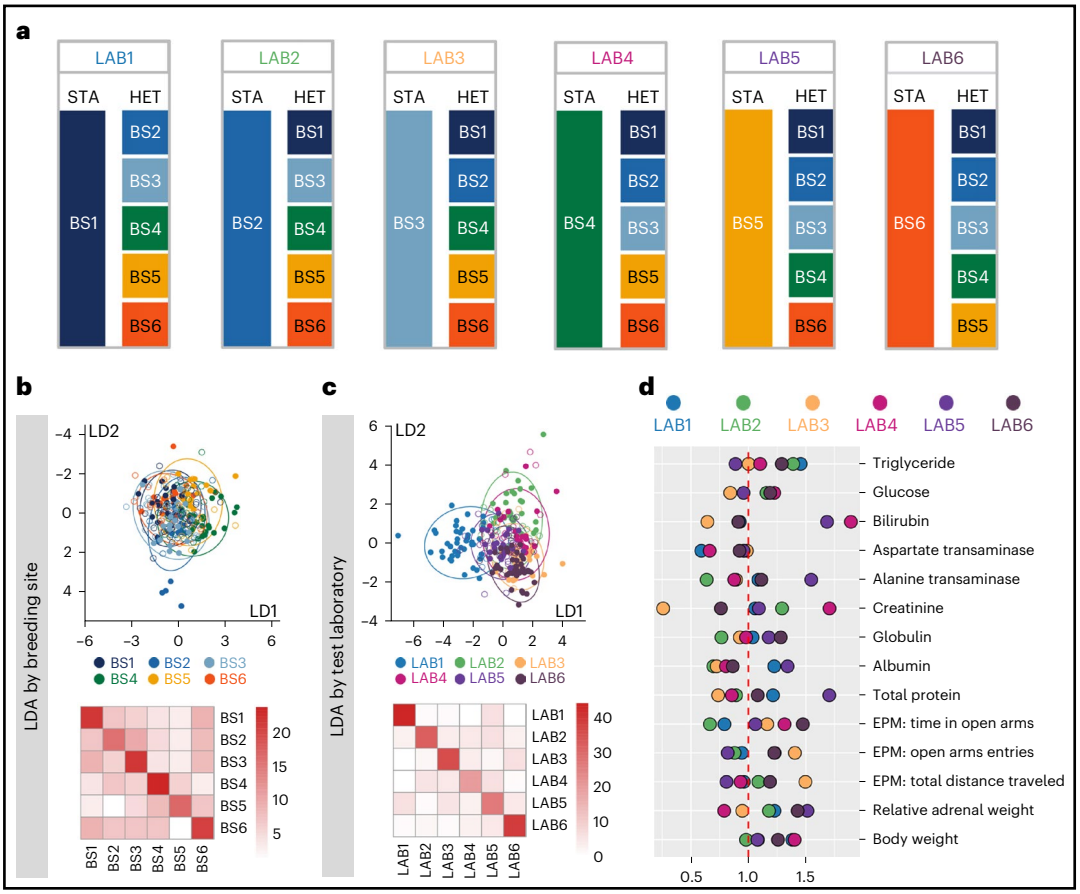

Fig. 1 | Effects of heterogenization on phenotypes. a, Schematic illustration of the multi-laboratory study design. **b**,**c**, The phenotype of mice was shaped by both breeding site (**b**) and test laboratory (**c**). In the LDA plots, color indicates breeder (**b**) and test laboratory (**c**). Filled dots represent a correct classification based on LDA, empty dots represent mismatches. **d**, Ratio of standard deviations of HET over STA cohorts for each outcome variable and each laboratory. A ratio >1 indicates that variation was larger in the HET compared to the STA cohort. BS, breeding site; LAB, test laboratory; LD, linear discriminant.

environmental background may serve as an effective heterogenization factor[14]. In this Article, we therefore tested whether systematic heterogenization of study populations, by using mice from different breeding sites to introduce the genetic and environmental variation that normally exists between independent study populations, would increase the external validity of the results sufficiently to guarantee replicability.

We used male C57BL/6J mice as a worked example and conducted a multi-laboratory study, with the same experiment conducted independently in six different laboratories by the same experimenter using the same test equipment. Each laboratory simultaneously employed both a standardized (STA) and a heterogenized (HET) study design (Fig. 1a). For STA, each laboratory ordered all mice ($n = 24$) in one cohort from one of six breeding sites (each laboratory from a different breeding site) to mimic the real-world situation of researchers independently ordering mice from a breeding site of their choice. By contrast, for HET, each laboratory ordered proportionate numbers of mice from the other five breeding sites ($n = 30$; 6 per lab; excluding the breeding site of the mice used in the STA design) to heterogenize the study population by the phenotypic variation that exists between mice from independent breeding sites (Fig. 1a). To test our hypothesis, we assessed 14 outcome variables, including body weight, behavioral measures obtained from a single session on an elevated plus maze (EPM), and clinical blood parameters. To eliminate potential sources of variation introduced by different experimenters and local test equipment, all mice underwent testing by the same experimenter using identical test equipment.

This approach allowed us to: (1) disentangle variance components originating from breeding site (combined effects of the genetic and environmental background) and test laboratory (where the experimental part of the study was performed); (2) test whether systematic heterogenization of study populations by using animals from different breeding sites increased the variance of the HET cohort compared to the STA cohort; and (3) evaluate the effectiveness of the HET design in improving replicability by meta-analyses for each outcome measure.

We found that both breeding site and test laboratory affected variation in outcome variables, but the effect of the test laboratory was much stronger than that of breeding site, despite the standardization of test equipment and test procedures. Since breeding site did not have a strong effect on variation in outcome variables, heterogenization by breeding site was not effective in improving the replicability of the results across laboratories.

Results

We obtained samples of 14 outcome variables from 308 mice, resulting in 4,283 outcome measures after accounting for missing values (Extended Data Table 1 and Methods). Both breeding site and test laboratory, as well as their interaction, had significant effects on variation in outcome variables (multivariate analysis of variance (MANOVA), Extended Data Table 2). Whereas laboratory explained 11.2% of the multivariate variance, breeding site accounted for only 4.0%, and 11.4% were due to the interaction between breeding site and test laboratory (η^2 estimates based on Pillai statistic). In a linear discriminant analysis (LDA) by breeding site, the first two discriminant functions explained 68% of the total variation (Extended Data Table 3), with LDA correctly predicting breeding site in 41% of cases (Fig. 1b) compared to 17% expected by chance. However, in

an LDA by test laboratory, the first two discriminant functions accounted for even 79% of the total variation and correctly predicted the test laboratory in 66% of cases (Fig. 1c and Extended Data Table 4).

Post-hoc analyses of variance for individual outcome measures with breeding site and test laboratory as fixed effects and cage as random effect confirmed that breeding site and test laboratory together explained on average 26% of the total variation (range 10–43%). Thus, both the origin of the animals (breeding site) and the test conditions (test laboratory) affected the outcome measures, but test laboratory had a stronger effect than breeding site. Indeed, in 13 out of 14 outcome variables, test laboratory accounted for more of the variance than breeding site (Extended Data Table 5). We note that in some cases the post-hoc models produced a singular or boundary fit, which means that the covariance matrix may not be estimated correctly. The outcomes of those analysis of variance (ANOVA) models should thus be interpreted carefully.

To assess whether heterogenization by breeding site increased within-laboratory variance at the expense of between-laboratory variance, we compared the variance of each outcome variable between STA and HET cohorts for each laboratory. Variance was larger in HET cohorts than in STA cohorts in 45 cases but smaller in 39 cases, although differences were generally small. A statistically significant difference between HET and STA cohorts was detected in only 1 out of the 84 contrasts (Levene tests for equal variances, Extended Data Table 6), which is even below the expected rate of false positive findings (4.2), given $\alpha = 0.05$. After adjusting the α-level threshold for multiple testing using a Bonferroni correction ($\alpha' = 5.9 \times 10^{-4}$), not a single statistically significant difference was detected. When combining outcome variables from the six test laboratories to obtain a single measure for each outcome variable, we did not find a significant difference between variances in the STA and HET cohorts for any of the 14 outcome measures (Extended Data Table 7). Thus, both at the level of individual contrasts and at the level of outcome variables, we found no evidence that variance was larger in HET cohorts.

Since each HET cohort contained animals from five of the six breeding sites used for STA cohorts, we expected lower between-laboratory variance in HET compared to STA cohorts. However, we found equal proportions of variance for test laboratory in HET cohorts (15.3%, range 4.9–40.4%) and STA cohorts (15.3%, range 3.7–40.3%). For 8 out of 14 outcome variables, between-laboratory variation was larger in STA cohorts, but for the other 6, it was larger in HET cohorts (Fig. 1d). We conclude that heterogenization of study populations by using mice from different breeding sites did not reduce between-laboratory variation.

Finally, treating the results from the six test laboratories as replicate studies, we conducted meta-analyses for each outcome variable for both the HET and STA study designs. We predicted that study means deviate less from the meta-analytic mean in HET cohorts than in STA cohorts. However, random effects meta-analyses showed similar results for both HET and STA cohorts (Fig. 2 and Supplementary Fig. 1). A mixed-effect model with cohort as fixed factor and outcome measure and test laboratory as random factors suggests that for all outcome variables, study design (HET versus STA) explained only 0.3% of the variation in the 'dance around the means' in the forest plots. On average, 76.2 ± 7.4% (mean ± standard deviation) of the study estimates for outcome measures from the HET cohorts fell within the 95% confidence interval of the meta-analytic mean estimate, compared to 73.8 ± 9.5% from the STA cohorts, providing no evidence for a higher coverage probability of estimates from HET cohorts compared to STA cohorts.

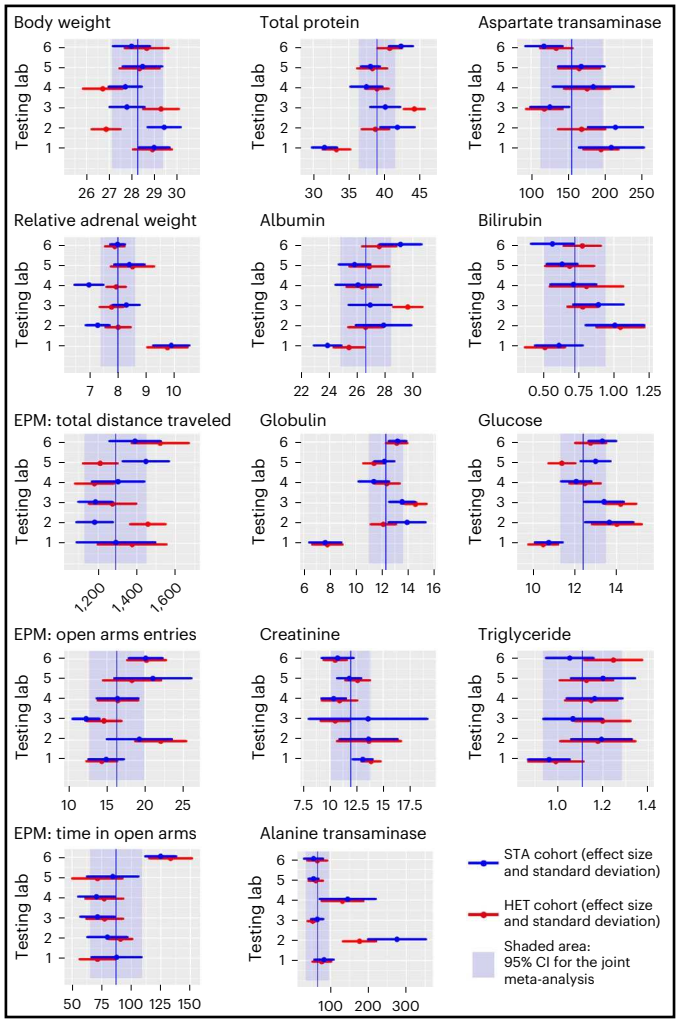

Fig. 2 | Meta-analysis for each outcome variable depending on study design. Heterogenization of study populations by breeding site (HET) did not improve replicability compared to a STA study design.

Discussion

The 'replicability crisis' in biomedical research calls for effective solutions[7,15]. Similar to multi-center trials in clinical research, multi-laboratory studies might be an ideal solution for preclinical animal studies, but their implementation can be challenging due to logistical demands and intellectual property concerns[10,14]. Effective heterogenization of study samples within single-laboratory studies could potentially be an alternative approach, mimicking the benefits of multi-laboratory studies without the logistical and intellectual property challenges[9].

Effective heterogenization of study populations requires the systematic variation of genetic and/or environmental factors that typically vary between independent replicate studies, thereby contributing to between-laboratory variation (that is, heterogeneity in meta-analyses) and thus poor replicability. Here we systematically tested whether heterogenization of study populations by including animals from different breeding sites is effective in improving the replicability of findings from single-laboratory animal studies.

The rationale behind choosing breeding sites as a heterogenization factor was based on our recent findings that common environmental differences between animal facilities can induce facility-specific phenotypes in mice[14]. Additionally, we considered the well-documented phenotypic variation that naturally occurs between different substrains of C57BL/6J mice[16–18]. Therefore, we expected the inclusion of mice from different breeding sites in study populations of single-laboratory studies to increase variation in many phenotypic traits, thereby mimicking the phenotypic variation that typically exists between different independent studies.

Contrary to our expectations, heterogenization of study populations by breeding site did not reduce between-laboratory variability compared to the conventional STA design. Several reasons may explain these unexpected findings. The main reason may be that breeding site contributed only little to total phenotypic variation, much less than test laboratory.

As a result, there was little scope for heterogenization by breeding site to reduce between-laboratory variation. Given our previous findings that common environmental differences between animal-rearing facilities can induce persistent phenotypic differences from the molecular to the behavioral level in mice[14], this finding was unexpected. One explanation could be that the rearing conditions in the facilities of professional breeders are much more similar (that is, STA) than the animal facilities of independent research institutions. Furthermore, the six breeding sites belonged to only three breeding companies. Thus, strictly STA operating procedures maintained across different breeding sites within companies could have further reduced phenotypic variation between mice from different breeding sites. Alternatively, the diversity of the mice within breeding sites may have been greater than expected, thereby limiting the scope for variation between breeding sites. This could, for example, be due to variation in age (the age of mice may vary by several days) and origin from different colony rooms.

The pair-housing of male mice could be another factor potentially contributing to larger diversity among mice from the same breeding site. Pair-housing may often result in despotic hierarchies among male mice, and it was found that circulating testosterone levels can differ by up to fivefold between dominant and subordinate males[19]. Such social effects may lead to substantial variability among mice of the same age and strain housed under identical conditions[20]. This may have been further corroborated by the need to single-house some animals for some time before testing due to escalating aggression. Although we accounted for this statistically, it remains possible that the biological effect was more pronounced[21,22].

Importantly, the effect of the test laboratory on phenotypic variability was considerably stronger than that of the breeding site. Previous studies[23–25] have indicated that the experimenter can have a strong influence on study outcomes, particularly emphasizing the impact of the experimenter's biological sex on behavioral outcomes in rodents. Despite deliberately harmonizing test procedures and equipment and having the same female experimenter conduct all test procedures in all six laboratories, the test laboratory still contributed substantially to the total variation in outcome variables. This suggests that other factors of housing and husbandry that varied between test laboratories (for example, cage ventilation, cage types, environmental enrichment and animal care) must have influenced outcome variables. Thus, laboratory-specific microenvironments may have shaped the phenotypic states of the mice, thereby influencing the study outcomes. Such effects of the test laboratory would normally be even stronger, as the test equipment and test procedures that were standardized in this study would normally vary between test laboratories[3,13,14].

In conclusion, we found no evidence that using mice from different breeding sites is potent enough to account for the variation that normally exists between results obtained in different laboratories. Although we here present a 'negative finding' or 'null-result', we believe that our study can serve as an example of how to implement heterogenization and how to assess the effectiveness of such an intervention on the external validity and replicability of experimental results. Our findings demonstrate substantial between-laboratory variation despite harmonized procedures, highlighting the need to strengthen our efforts to find practicable ways of heterogenizing study populations effectively to improve the replicability of results from basic and preclinical animal research.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41684-023-01307-w.

Received: 4 May 2023; Accepted: 20 November 2023
Published online: 27 December 2023

References

1. Beynen, A. C., Gärtner, K. & van Zutphen, L. F. M. in *Principles of Laboratory Animal Science* Ch. 5 (eds. Zutphen, L. F. M., Baumans, V. & Beynen, A. C.) 103–110 (Elsevier, 2001).
2. Crabbe, J. C., Wahlsten, D. & Dudek, B. C. Genetics of mouse behavior: interactions with laboratory environment. *Science* **284**, 1670–1672 (1999).
3. Chesler, E. J., Wilson, S. G., Lariviere, W. R., Rodriguez-Zas, S. L. & Mogil, J. S. Influences of laboratory environment on behavior. *Nat. Neurosci.* **5**, 1101–1102 (2002).
4. Corrigan, J. K. et al. A big-data approach to understanding metabolic rate and response to obesity in laboratory mice. *eLife* **9**, e53560 (2020).
5. Würbel, H. Behaviour and the standardization fallacy. *Nat. Genet.* **26**, 263 (2000).
6. Richter, S. H., Garner, J. P. & Würbel, H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? *Nat. Methods* **6**, 257–261 (2009).
7. Voelkl, B. et al. Reproducibility of animal research in light of biological variation. *Nat. Rev. Neurosci.* **21**, 384–393 (2020).
8. Richter, S. H. et al. Effect of population heterogenization on the reproducibility of mouse behavior: a multi-laboratory study. *PLoS ONE* **6**, e16461 (2011).
9. Richter, S. H., Garner, J. P., Auer, C., Kunert, J. & Würbel, H. Systematic variation improves reproducibility of animal experiments. *Nat. Methods* **7**, 167–168 (2010).
10. Voelkl, B., Vogt, L., Sena, E. S. & Würbel, H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. *PLoS Biol.* **16**, e2003693 (2018).
11. Voelkl, B. & Würbel, H. A reaction norm perspective on reproducibility. *Theory Biosci.* **140**, 169–176 (2021).
12. Bailoo, J. D. et al. Effects of weaning age and housing conditions on phenotypic differences in mice. *Sci Rep.* **10**, 11684 (2020).
13. Von Kortzfleisch, V. T. et al. Do multiple experimenters improve the reproducibility of animal studies? *PLoS Biol.* **20**, e3001564 (2022).
14. Jaric, I. et al. The rearing environment persistently modulates mouse phenotypes from the molecular to the behavioural level. *PLoS Biol.* **20**, e3001837 (2022).
15. Munafò, M. R. et al. A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).
16. Mekada, K. et al. Genetic differences among C57BL/6 substrains. *Exp. Anim.* **58**, 141–149 (2009).
17. Mekada, K., Hirose, M., Murakami, A. & Yoshiki, A. Development of SNP markers for C57BL/6N-derived mouse inbred strains. *Exp. Anim.* **64**, 91–100 (2015).
18. Mekada, K. & Yoshiki, A. Substrains matter in phenotyping of C57BL/6 mice. *Exp. Anim.* **70**, 145–160 (2021).
19. Machida, T., Yonezawa, Y. & Noumura, T. Age-associated changes in plasma testosterone levels in male mice and their relation to social dominance or subordinance. *Horm. Behav.* **15**, 238–245 (1981).
20. Varholick, J. A. et al. Social dominance hierarchy type and rank contribute to phenotypic variation within cages of laboratory mice. *Sci Rep.* **9**, 13650 (2019).
21. Arndt, S. S. et al. Individual housing of mice—impact on behaviour and stress responses. *Physiol. Behav.* **97**, 385–393 (2009).
22. Bartolomucci, A. et al. Individual housing induces altered immuno-endocrine responses to psychological stress in male mice. *Psychoneuroendocrinology* **28**, 540–558 (2003).
23. Sorge, R. E. et al. Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nat. Methods* **11**, 629–632 (2014).
24. Mogil, J. S. Laboratory environmental factors and pain behavior: the relevance of unknown unknowns to reproducibility and translation. *Lab Anim.* **46**, 136–141 (2017).

25. Georgiou, P. et al. Experimenters' sex modulates mouse behaviors and neural responses to ketamine via corticotropin releasing factor. *Nat. Neurosci.* **25**, 1191–1200 (2022).

Methods

Ethical statement

All animal experiments were conducted in full compliance with the Swiss Animal Welfare Ordinance (TSchV 455.1) and were approved by the Cantonal Veterinary Office in Bern, Switzerland (permit number BE88/20).

Animal subjects and study design

In this multi-laboratory study, we focused on the C57BL/6J strain, as it is the most widely used strain in biomedical research[26–28]. As this was a proof-of-principle study, and to keep the study manageable, only male subjects were used. We selected male mice on the basis of our recent work, which demonstrated more pronounced phenotypic differences in C57BL/6J males raised in different facilities[14].

In this study, we investigated the effectiveness of using animals from multiple breeding sites to introduce genetic and environmental variation as a solution to systematically increase variation within a single test laboratory and, consequently, decrease variation between test laboratories.

Mice were obtained from the following six commercial breeding sites (Supplementary Fig. 2):

i. Charles River Laboratories DE, Sulzfeld, Germany (B1; C57BL/6JCrl mice);
ii. Charles River Laboratories FRA, L'Arbresle, France (B2; C57BL/6JCrl mice);
iii. Charles River Laboratories UK, Kent, United Kingdom (B3; C57BL/6JCrl mice);
iv. Envigo RMS, Gannat, France (B4; C57BL/6JOlaHsd mice);
v. Envigo RMS, Gannat, France (B5; C57BL/6JRccHsd mice);
vi. Janvier Labs, Le Genest-Saint-Isle, France (B6; C57BL/6JRj mice).

The test laboratories were located at the following institutions:

i. Institute of Anatomy, University of Zürich (LAB 1);
ii. Division of Animal Welfare, Vetsuisse Faculty, University of Bern (LAB 2 and LAB 4);
iii. Central Animal Facilities, Experimental Animal Center, University of Bern (LAB 3);
iv. Institute of Veterinary Pharmacology and Toxicology, Vetsuisse Faculty, University of Zürich (LAB 5);
v. Laboratory of Neuroepigenetics, Brain Research Institute, University of Zurich and Institute for Neuroscience, ETH Zurich (LAB 6).

Each test laboratory provided space for animal housing, a test room for behavioral testing and an experimental room for tissue collection. Animal care was provided by each laboratory's animal care staff.

For the STA study design, each test laboratory ordered all mice in one cohort from one of six breeding sites (each laboratory from a different site). For the systematically HET study design, each test laboratory received mice in proportionate numbers from five of the six breeding sites (excluding the one from which they ordered the mice for the STA study design). This resulted in a total of 12 replicate experiments (6 STA and 6 HET), in which 324 mice were used. The final number of mice used for STA and HET design in each test laboratory is presented in the data file.

As each test laboratory conducted the experiment independently, animals were delivered separately at an age of 12 weeks ($n = 54$ per test laboratory). The mice were shipped in groups of two cagemates in small or subdivided boxes. Due to the predisposition to elevated aggressive behavior in C57BL/6J males, the animals shipped together were housed together upon arrival.

Upon arrival, the animals were checked for health, then individually marked by fur cut, randomly assigned to cages by breeding site, and pair-housed under laboratory-specific housing and husbandry conditions (Supplementary Table 1) for 12 days before the onset of the test phase (Supplementary Fig. 3). Cage positions on the rack were also counterbalanced by breeding site (animal origin) and study design (STA or HET). Cages were cleaned 7 days after arrival and left undisturbed until the onset of the test phase to minimize disruption due to cage cleaning before testing. Food pellets and tap water were provided ad libitum. All mice were held under a constant 12-h light–dark cycle, but the time schedules differed between laboratories (Supplementary Table 1).

Since it has been shown that the test environment can have a profound influence on study outcomes[2,3,29], the effects are very often attributed to the differences in test protocols (test time, equipment, illumination and so on)[30] that normally exist between different laboratories. Thus, we controlled for all those factors by standardizing the test protocol and equipment across all six test laboratories. Additionally, studies have suggested that the experimenter performing the tests might have an effect on the outcome measures[23–25], and that effect might be even stronger than the effect of the genotype on the same outcome measure[3]. In our experimental setup, we wanted to exclude that possibility, so the same experimenter (I.J.) performed behavioral testing and tissue collection in each test laboratory, thereby minimizing procedural variation that might affect outcome measures.

Sample size calculation

The sample size for the HET study was partly determined by the requirement for a balanced study design within the HET cohorts. The sample size for the STA design was then incrementally adjusted until an estimated power of 0.8 was reached. To estimate the achieved power, we used simulated sampling. The R code for this simulation is attached as a supplementary file. In short, following simulated sampling with specific assumptions for the distribution of expected effect sizes, a principal component analysis was conducted over all 12 variables using orthogonal rotation, and the first principal component was taken as the input for an ANOVA analysis. The analysis aimed to determine how often the f ratios of the means squares for the HET and STA designs exceeded the threshold value of $f = 6.6$ ($P \leq 0.05$ for 1 and 5 d.f.). The results showed that, under these assumptions, a significant main effect was found in 82.5% of the cases for a sample size of 24 animals in the STA cohort, indicating an achieved power of 0.825.

Behavioral testing

To analyze phenotypic variation in behavior, we focused on changes in exploratory and emotional behavior by using one of the most common behavioral assay, the EPM[31,32].

EPM testing was carried out in batches over two consecutive days during the dark phase, specifically between the first and fourth hours. The EPM apparatus was made of a gray-colored polycarbonate platform with a white removable floor. The platform comprised two opposite open arms (30 cm × 6 cm) and two opposite closed arms surrounded by 15-cm-high walls of the same dimensions. The central part that allows the animal to transit from arm to arm consists of a square with dimensions of 6 × 6 cm. The maze was elevated 40 cm above the ground, and the open arms were equipped with a small lip around the perimeter, 0.5 cm high, to ensure that no animals would fall off the maze. The illumination at the open arms was set to 140 lux.

Each test started by taking the mouse from the home cage and placing it in the center part of the EPM, facing the closed arm. Mice were allowed to freely explore the maze for 5 min. Both cagemates were tested simultaneously using two identical apparatuses placed next to each other but visually separated. The test order was balanced across breeding sites and experimental designs and randomized using the random number generator of the Mathematica software (version 11; Wolfram Research) separately for each test laboratory. Between trials, the apparatuses were

sprayed with water containing odorless detergent, rinsed two times with water, and dried with paper towels.

The total distance traveled, the time spent in the open arms, and the number of entries into the open arms were measured from video recordings using EthoVision XT software (version 11.5; Noldus). The criterion for arm entry was when the center point of the animal (as detected by Ethovision) was in the arm.

Tissue sampling procedure

Two days after the EPM test, animals were weighed and deeply anesthetized with an overdose of pentobarbital diluted in 0.9% saline (150 mg/kg, Esconarkon, Streuli Pharma AG). To avoid possible influences of the circadian rhythm on the blood clinical parameters, the procedures were performed during the first four hours of the light phase. The order of trials corresponds to the one used for behavioral testing.

Approximately 600–800 µl of blood was collected by cardiac puncture and transferred into potassium ethylenediaminetetraacetic acid (EDTA)-coated tubes (Micro sample tube K3 EDTA, 1.6 mg EDTA/ml blood, Sarstedt). Immediately after the puncture, the blood samples were placed on ice, and the animals were decapitated. Adrenal glands were removed, dissected from fat, and weighed using a precision scale (Mettler AE160, Mettler-Toledo). Within 1 h, the blood samples were centrifuged for 10 min at 4,000*g* and 4 °C. Plasma samples were transferred to new, labeled microcentrifuge tubes and stored at −80 °C until assayed.

Blood clinical chemistry

We focused on blood chemistry parameters since they provide a good overview of the metabolic state and organ functions, as well as electrolyte and mineral homeostasis[33].

All analyses were performed on a Roche Cobas c501 analyzer (Roche Diagnostics (Schweiz) AG). Total protein, albumin, globulin, creatinine, triglycerides and glucose, as well as the enzymatic activity of alanine transaminase and aspartate aminotransferase, were quantified photometrically with reagents provided by Roche Diagnostics. All procedures were performed according to the manufacturers' protocols.

Statistical analysis

Statistical analyses were performed in R (Supplementary Code)[34]. All analyses were performed for the same set of outcome variables: body weight at the day of sacrificing, relative adrenal weight, total distance traveled in the EPM, number of open arm entries in the EPM, time spent in the open arms in the EPM, and the blood plasma concentration of total protein, albumin, globulin, creatinine, enzymatic activity of alanine transaminase and aspartate aminotransferase, bilirubin, glucose and triglyceride.

To identify variance components attributable to breeder and laboratory, we first made a MANOVA with the outcome measures as dependent variables and laboratory, breeder and the interaction between laboratory and breeder as independent variables, followed by post-hoc mixed-effect regression models for each outcome variable with breeder, laboratory and their interaction as fixed factors and cage ID as a random factor. For calculating *P* values for the mixed-effect regression models, degrees of freedom were estimated using the Satterthwaite approximation.

Following the MANOVA, two separate LDAs were made: one with breeder as the response variable and the outcome variables as linear predictors, and one with laboratory as a response variable.

For comparing the variance of each outcome variable between STA and HET cohorts within each of the laboratory, we used Levene tests for equal variances with the significance threshold set to $\alpha = 0.05$ (without correction for multiple testing). For combining outcome variables from the six laboratories to obtain a single measure for each outcome variable, we used Fisher's method for combined probabilities[35].

To investigate whether HET designs led to lower between-laboratory variation than STA designs, we ran for each outcome variable two separate mixed models with the outcome as a dependent variable, laboratory as fixed effect and cage ID as a random factor—one for the HET design and one for the STA design. We then compared the marginal R^2 estimates. In a final step, we treated the results from the six laboratories as replicate studies and conducted random effect meta-analyses[36] with the outcome as dependent variable and laboratory as random effect.

Blinding

The experimenter performing weighing, the EPM test and tissue collection was blind to the 'study design', that is, STA or HET design. Blinding was done by two colleagues otherwise not involved in the execution of the experiments. Cages were assigned identification numbers so that the experimenter could not deduce the origin of the cages (that is, breeding site) from the ID number or the position of the cage. Blinding with regard to the test laboratory was not possible for weighing and organ collection since the experimenter needed to travel to each test facility. For the clinical chemistry analysis, the experimenter was blind to the study design and the test laboratory as well.

Missing data and cases of single housing

During the experiment, a total of 16 mice were lost. In test laboratory 1, one mouse was euthanized immediately after arrival due to poor health conditions. The animal was apathetic and cold, and had wounds upon arrival. As a result, its cagemate was housed alone for the whole duration of the study. Additionally, in two cages, we observed an increased incidence of fighting, which resulted in small bite wounds. Consequently, a total of four animals had to be housed separately for a short period before behavioral testing.

In test laboratory 2, there was no need for single housing, and no animals were lost.

In test laboratory 3, three animals were euthanized in consultation with the responsible veterinarian due to a high level of wounding. This occurred 2 days before testing, resulting in the brief single housing of their cagemates.

In test laboratory 4, two cagemates were found dead during the habituation period; however, necropsy did not reveal a specific cause of death. Additionally, two more mice were euthanized due to high levels of injuries that occurred between two daily checks. Consequently, their cagemates were also single-housed. Furthermore, due to an observed incidence of aggression with tail bites, six additional mice from three cages were single-housed 2 days before testing.

In test laboratory 5, a total of five animals were lost. Cagemates from two cages, a total of four mice, were lost during the habituation period, while one animal was lost just before tissue collection, which did not result in the single housing of its cagemate. The necropsy of that animal showed the presence of cysts on both kidneys. Furthermore, 14 mice from seven cages needed to be individually housed due to incidences of aggression.

In test laboratory 6, one animal was euthanized immediately after arrival due to poor health conditions. Its cagemate was single-housed for the whole duration of the experiment. Additionally, two more animals were found dead in the home cage, 3 days before behavioral testing, which resulted in a period of single housing for their cagemates. Moreover, in three cages, a total of six mice needed to be single-housed 4 days before testing until the end of the experiment.

For the EPM testing, 25 data points were lost for each EPM outcome measure. Fourteen mice had their data lost due to animal euthanasia or death before testing, and an additional nine data points were lost due to technical problems during the transfer of recorded videos.

Two additional data points for blood clinical chemistry were excluded due to measurement errors.

Recalculating power given the final number of animals entering the analysis (299 for behavioral measures), there was a small drop in statistical power from 82.5% to 79.9% (Supplementary Material).

Reporting summary

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

Data availability

All data supporting the findings of this study together with the code are available within the article and its supplementary files (Supplementary Data and Supplementary Code).

References

26. Fontaine, D. A. & Davis, D. B. Attention to background strain is essential for metabolic research: C57BL/6 and the International Knockout Mouse Consortium. *Diabetes* **65**, 25–33 (2016).
27. Marchette, R. C. N., Bicca, M. A., Santos, E. C., da, S. & de Lima, T. C. M. Distinctive stress sensitivity and anxiety-like behavior in female mice: strain differences matter. *Neurobiol. Stress* **9**, 55–63 (2018).
28. Bryant, C. D. The blessings and curses of C57BL/6 substrains in mouse genetic studies: Bryant. *Ann. N. Y. Acad. Sci.* **1245**, 31–33 (2011).
29. Butler-Struben, H. M., Kentner, A. C. & Trainor, B. C. What's wrong with my experiment?: the impact of hidden variables on neuropsychopharmacology research. *Neuropsychopharmacology* **47**, 1285–1291 (2022).
30. Saré, R. M., Lemons, A. & Smith, C. B. Behavior testing in rodents: highlighting potential confounds affecting variability and reproducibility. *Brain Sci.* **11**, 522 (2021).
31. Rosso, M. et al. Reliability of common mouse behavioural tests of anxiety: a systematic review and meta-analysis on the effects of anxiolytics. *Neurosci. Biobehav. Rev.* **143**, 104928 (2022).
32. Pawlak, C. R., Karrenbauer, B. D., Schneider, P. & Ho, Y.-J. The elevated plus-maze test: differential psychopharmacology of anxiety-related behavior. *Emot. Rev.* **4**, 98–115 (2012).
33. Mouse Phenome Database Team. et al. A comprehensive and comparative phenotypic analysis of the collaborative founder strains identifies new and known phenotypes. *Mamm. Genome* **31**, 30–48 (2020).
34. R Core Team. R: a language and environment for statistical computing. *R Foundation for Statistical Computing* https://www.R-project.org/ (2021).
35. Sokal, R. R. & Rohlf, F. J. *Biometry: The Principles and Practice of Statistics in Biological Research* 3rd Edn (W.H. Freeman and Co., 1995).
36. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36**, 1–48 (2010).

Acknowledgements

This study was supported by the Swiss National Science Foundation (grant 310030_179254) to H.W.

Author contributions

I.J., B.V. and H.W. designed the study. I.J. coordinated and directed the project, conducted behavioral testing and analysis, and collected and analyzed tissue samples. B.V. performed statistical analysis of phenotypic measurements with the input from I.J. and H.W. I.A., J.N., U.W.-S. and F.M. assisted with tissue collection within test facilities. C.D., D.P.W., U.M. and I.M.M. provided the facility resources and laboratory space. I.J., B.V. and H.W. interpreted the data. I.J. and B.V. constructed the figures. H.W. provided the main funding and supervised the project. I.J., B.V. and H.W. wrote the manuscript with input from all authors.

Funding

Open access funding provided by University of Bern.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41684-023-01307-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41684-023-01307-w.

Correspondence and requests for materials should be addressed to Ivana Jaric or Hanno Würbel.

Peer review information *Lab Animal* thanks Ulf Tölch and other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Extended Data Table 1 | Number of recorded values per outcome measure

Phenotype measure	Number of recorded values
Body weight	308
Relative adrenal weight	308
Elevated Plus Maze: total distance travelled	299
Elevated Plus Maze: open arms entries	299
Elevated Plus Maze: time in open arms	299
Total protein	308
Albumin	308
Globulin	308
Creatinine	306
Alanine transaminase	308
Aspartate transaminase	308
Bilirubin	308
Glucose	308
Triglyceride	308

Extended Data Table 2 | MANOVA outcome

Factor	df	Pillai	approx F	df_{num}	df_{den}	$p(>F)$
LAB	5	1.572	8.190	70	1250	$< 2.2\times10^{-16}$
BREEDER	5	0.562	2.262	70	1250	3.885×10^{-08}
LAB×BREEDER	25	1.591	1.328	350	3626	8.966×10^{-05}
Residuals	259					

Model formula: Y ~ LAB + BREEDER + LAB×BREEDER. LAB: Test laboratory; BREEDER: Breeding site; df: degrees of freedom of the dependent variables, Pillai: Pillai's Trace statistic; df_{den}: denominator degrees of freedom based on Satterthwaite's approximation.

Extended Data Table 3 | Linear discriminant function analysis for breeding site

Phenotype measure	LD1	LD2	LD3	LD4
Body weight	0.126	-0.040	-0.029	0.067
Relative adrenal weight	-0.339	-0.021	-0.110	-0.081
Elevated Plus Maze: total distance travelled	-7.74×10^{-4}	3.72×10^{-4}	9.35×10^{-06}	0.001
Elevated Plus Maze: open arms entries	-0.023	0.045	0.021	0.022
Elevated Plus Maze: time in open arms	-0.003	0.003	0.011	-0.005
Total protein	0.064	0.424	0.230	-0.335
Albumin	0.041	-0.467	-0.025	0.240
Globulin	-0.011	-0.460	-0.096	0.574
Creatinine	-0.006	-0.076	-0.038	0.035
Alanine transaminase	0.005	-5.93×10^{-04}	3.70×10^{-04}	-0.001
Aspartate transaminase	0.002	0.005	-3.30×10^{-05}	-2.92×10^{-04}
Bilirubin	-0.121	0.097	-0.707	0.027
Glucose	-0.268	-0.160	-0.124	0.003
Triglyceride	0.713	1.856	-1.677	0.869

LD1-4: Coefficients of the linear discriminants displaying the linear combination of predictor variables.

Extended Data Table 4 | Linear discriminant function analysis for laboratory

Phenotype measure	LD1	LD2	LD3	LD4
Body weight	-0.244	-0.114	0.125	0.215
Relative adrenal weight	-0.517	-0.130	0.045	0.267
Elevated Plus Maze: total distance travelled	-6.76×10^{-06}	-2.42×10^{-04}	9.87×10^{-04}	-6.90×10^{-05}
Elevated Plus Maze: open arms entries	-3.26×10^{-03}	0.023	0.033	-0.021
Elevated Plus Maze: time in open arms	-8.86×10^{-04}	-0.009	0.015	-4.59×10^{-03}
Total protein	0.080	0.129	-0.026	-0.484
Albumin	-0.089	-0.227	0.135	0.474
Globulin	0.261	-0.239	-0.091	0.431
Creatinine	-0.031	0.011	0.008	0.041
Alanine transaminase	5.42×10^{-04}	0.006	0.003	6.19×10^{-04}
Aspartate transaminase	-2.81×10^{-03}	0.003	5.72×10^{-04}	-2.45×10^{-03}
Bilirubin	0.357	0.816	0.131	0.782
Glucose	-0.017	0.071	0.080	0.364
Triglyceride	-0.532	0.548	-0.128	-0.662

LD1-4: Coefficients of the linear discriminants displaying the linear combination of predictor variables.

Extended Data Table 5 | Outcomes for post-hoc type III ANOVAs

Phenotype measure	$F_{BREEDER}$ (df_{num},df_{den})	F_{LAB} (df_{num},df_{den})	$F_{BREEDERxLAB}$ (df_{num},df_{den})	$R^2_{marginal}$	$R^2_{conditional}$
Body weight	5.104 (5,124.78)	8.032 (5,124.76)	4.224 (25,123.88)	0.373	0.426
Relative adrenal weight	4.016 (5,121.85)	11.891 (5,121.85)	1.136 (25,121.14)	0.356	0.514
Elevated Plus Maze: total distance travelled	3.128 (5,263)	3.882 (5,263)	1.475 (25,263)	0.208	0.208
Elevated Plus Maze: open arms entries	4.186 (5,263)	6.171 (5,263)	1.072 (25,263)	0.232	0.232
Elevated Plus Maze: time in open arms	2.104 (5,128.90)	10.040 (5,127.90)	1.422 (25,124.95)	0.277	0.340
Total protein	1.496 (5,119.83)	18.990 (5,119.79)	0.933 (25,118.28)	0.361	0.419
Albumin	1.304 (5,122.38)	5.947 (5,122.34)	0.883 (25,121.34)	0.198	0.263
Globulin	1.526 (5,121.71)	27.203 (5,121.68)	1.110 (25,120.72)	0.433	0.491
Creatinine	1.112 (5,121.18)	1.876 (5,121.20)	0.410 (25,120.61)	0.102	0.347
Alanine transaminase	0.629 (5,272)	11.224 (5,272)	0.801 (25,272)	0.300	0.300
Aspartate transaminase	0.736 (5,125.09)	4.787 (5,125.02)	0.745 (25,123.93)	0.169	0.195
Bilirubin	0.426 (5,128.22)	5.494 (5,128.18)	1.527 (25,127.18)	0.207	0.272
Glucose	0.326 (5,125.41)	14.426 (5,125.37)	1.270 (25,124.41)	0.288	0.361
Triglyceride	3.457 (5,127.17)	2.553 (5,127.10)	1.499 (25,126.00)	0.187	0.211

In all cases mixed-effect models of Y ~ LAB + BREEDER + LAB×BREEDER + (1|Cage) were used. LAB: Test laboratory; BREEDER: Breeding site; df_{num}: numerator degrees of freedom; df_{den}: denominator degrees of freedom based on Satterthwaite's approximation. $R^2_{marginal}$: variance component of fixed effects, $R^2_{conditional}$: variance component of fixed and random effects.

Extended Data Table 6 | Levene Tests for equal variances

	F_{Lab1}	F_{Lab2}	F_{Lab3}	F_{Lab4}	F_{Lab5}	F_{Lab6}	$p_{Combined}$
Bodyweight	2.092	0.035	0.131	4.008	0.372	0.970	0.294
Relative adrenal weight	1.333	0.016	0.051	0.032	1.410	1.559	0.656
Elevated Plus Maze: total distance travelled	0.384	0.124	3.508	0.113	0.397	0.485	0.565
Elevated Plus Maze: open arms entries	0.001	0.029	2.557	0.089	1.156	0.386	0.710
Elevated Plus Maze: time in open arms	0.151	2.270	1.136	0.488	0.002	1.543	0.473
Total protein	0.006	0.110	4.261	0.685	3.808	0.159	0.231
Albumin	0.012	1.134	3.734	0.694	0.973	0.934	0.271
Globulin	0.013	0.131	0.236	0.076	0.134	2.588	0.839
Creatinine	1.68×10^{-04}	8.13×10^{-04}	2.442	0.210	0.190	0.008	0.913
Alanine transaminase	0.046	0.608	0.250	0.028	1.946	0.026	0.853
Aspartate transaminase	2.170	0.264	0.085	1.060	0.066	0.244	0.695
Bilirubin	0.399	0.012	1.208	0.579	1.665	0.009	0.707
Glucose	0.102	0.401	0.543	1.171	0.074	0.668	0.776
Triglyceride	1.315	2.729	0.339	0.036	0.031	0.986	0.508

F: F-statistic for the Levene test with 1 degree of freedom; $p_{Combined}$: combined probability over all 6 laboratories based on Fisher's method for combined probabilities

Extended Data Table 7 | Variance explained by the factor Laboratory

Phenotype measure	R^2_{HET}	R^2_{STA}
Body weight	0.163	0.101
Relative adrenal weight	0.196	0.379
Elevated Plus Maze: total distance travelled	0.117	0.087
Elevated Plus Maze: open arms entries	0.127	0.137
Elevated Plus Maze: time in open arms	0.198	0.160
Total protein	0.313	0.370
Albumin	0.143	0.172
Globulin	0.404	0.403
Creatinine	0.087	0.037
Alanine transaminase	0.195	0.342
Aspartate transaminase	0.108	0.148
Bilirubin	0.109	0.136
Glucose	0.263	0.199
Triglyceride	0.050	0.079

For all phenotype measures mixed-effect models of Y ~ LAB + (1|Cage) were built separately for the heterogenized (HET) and standardized (STA) study cohorts and the marginal R^2 estimates for the fixed factor lab (test laboratory) were estimated.

nature portfolio

Corresponding author(s): Ivana Jaric; ivana.jaric@unibe.ch
Hanno Würbel; hanno.wuerbel@unibe.ch

Last updated by author(s): Nov 14, 2023

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies 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	Behavioral data were collected from video recordings using EthoVision XT software (version 11.5; Noldus, Wageningen, the Netherlands). Adrenal glands data were generated by using a precision scale (Mettler AE160, Mettler-Toledo, Switzerland). Blood clinical chemistry analyses were performed on a Roche Cobas c501 analyzer (Roche Diagnostics (Schweiz) AG, Rotkreuz, Switzerland).
Data analysis	Statistical analyses are performed in R and script is provided as supplementary code and sours data files.

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Provide your data availability statement here.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

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

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 sample size for the heterogenized (HET) study was partly determined by the requirement for a balanced study design within the HET cohorts. The sample size for the standardized (STA) design was then incrementally adjusted until an estimated power of 0.8 was reached. In order to estimate the achieved power, we used simulated sampling. The R-code for this simulation is attached as a supplementary file. In short, following simulated sampling with specific assumptions for the distribution of expected effect sizes, a principal component analysis was conducted over all 12 variables using orthogonal rotation, and the first principal component was taken as the input for an ANOVA analysis. The analysis aimed to determine how often the f-ratios of the means squares for the HET and STA designs exceeded the threshold value of f=6.6 (p<=0.05 for 1 and 5 df). The results showed that under these assumptions, a significant main effect was found in 82.5% of the cases for a sample size of 24 animals in the STA cohort, indicating an achieved power of 0.825.
Data exclusions	During the experiment, a total of 16 mice were lost. In testing laboratories 1 and 6, two mice were euthanized immediately after arrival due to poor health conditions. In testing laboratory 4, two mice were found dead during the habituation period, while in testing laboratories 3 and 5, two mice were found dead just before the final tissue collection. However, necropsy did not reveal a specific cause of death. Additionally, during the habituation period, 11 mice were euthanized due to high levels of wounding: two in testing laboratory 3, two in testing laboratory 4, four in testing laboratory 5, and two in testing laboratory 6. All euthanasia procedures were performed after consultation with the responsible veterinarians in each testing facility. For the EPM testing, 25 data points were lost for each EPM outcome measure. Fourteen mice had their data lost due to animal euthanasia or death prior to testing, and an additional nine data points were lost due to technical problems during the transfer of recorded videos. Two additional data points for blood clinical chemistry were excluded due to measurement error
Replication	Since this was animal study, we used biological replicates.
Randomization	The animals were randomly assigned to cages by breeding site. Cage positions on the rack were also counterbalanced by breeding site (animal origin) and study design (STA or HET).
Blinding	The experimenter performing weighing, EPM test and tissue collection was blind to the "study design", i.e STA or HET design. Blinding was done by two colleagues otherwise not involved in the execution of the experiments. Cages were assigned identification numbers so that the experimenter cannot deduce the origin of the cages (i.e. breeding site) from the ID number or the position of the cage. Blinding with regard to testing laboratory was not possible for weighing and organ collection since the experimenter needed to travel to each testing facility. For the clinical chemistry analysis, the experimenter was blind to the "study design" and the testing laboratory, as well.

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 and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	We focused on the C57BL/6J male mice (12 week old) which were obtained from multiple breeding sites to introduce genetic and environmental variation. Mice were obtained from the following six commercial breeding sites: i) Charles River Laboratories DE, Sulzfeld, Germany (B1; C57BL/6JCrl mice); ii) Charles River Laboratories FRA, L'Arbresle, France (B2; C57BL/6JCrl mice); iii) Charles River Laboratories UK, Kent, United Kingdom (B3; C57BL/6JCrl mice); iv) Envigo RMS, Gannat, France (B4; C57BL/6JOlaHsd mice); v) Envigo RMS, Gannat, France (B5; C57BL/6JRccHsd mice); vi) Janvier Labs, Le Genest-Saint-Isle, France (B6; C57BL/6JRj mice).
Wild animals	This study did not include wild-animals.
Reporting on sex	As this was a proof-of-principle study and to keep the study manageable, only male subjects were used. We selected male mice based on our recent work, which demonstrated more pronounced phenotypic differences in C57BL/6J males raised in different facilities (Jaric et al. 2022 doi.org/10.1371/journal.pbio.3001837).
Field-collected samples	This study did not include field-collected samples.
Ethics oversight	All animal experiments were conducted in full compliance with the Swiss Animal Welfare Ordinance (TSchV 455.1) and were approved by the Cantonal Veterinary Office in Bern, Switzerland (permit number: BE88/20).

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