**S2 Text: Details on data analysis**

**Reproducibility of the strain effect across laboratories**

I.) Consistency of the strain effect across laboratories

The following linear mixed model (LMM, equation 1) was applied to both designs (standardised and heterogenised) to assess the consistency of the strain effect across laboratories (‘strain-by-laboratory’-interaction) as a measurement of reproducibility.

Model written in formal terms:

*yijmk* = *µ* + *ai* + *bj* + cij + *dm* + *fim* + *εijmk* Equation (1)

where *i* = 1, …, *nS*, *j* = 1, …, *nR*, *m* = 1,…, *nB* and *k* = 1,…, *nijm*. *ai* indicates the main effect of the *i*th level of strain (treatment); *bj* represents laboratory as a random effect where *bj*~ N(0,σb2); *cij* represents strain-by-laboratory-interaction as random effect where *cij*~ N(0,σc2); *dm* represents block as a random effect nested in laboratory where *dm*~ N(0,σd2); *fim* represents strain-by-block-interaction as a random effect where *fim*~ N(0,σf2) and the error term *εijmk* ∼ N(0,σe2).

Equation 1 written in layperson terms:

y = ‘strain’ + ‘laboratory’ + ‘strain x laboratory’ + ‘block’ + ‘block x strain’ Equation (1)

where ‘strain’ was included as fixed factor and ‘laboratory’, ‘strain-by-laboratory’-interaction, ‘block’ and ‘block-by-strain’-interaction as random factors. Moreover, the factor ‘block’ was nested within ‘laboratory’. The blocking factor was included in accordance with the randomised block design used, in which mice sharing the same micro-environment were treated as one ‘block’ (i.e. same rack position, see Fig 2). To check for the assumptions of parametric analysis, residuals were graphically examined for normal distribution, homoscedasticity, and the Shapiro-Wilk test was applied. When necessary, raw data were transformed using square root, logarithmic or arcsine transformations (see Supplementary S4 Table).

Typically, the contribution of a fixed effect to a model is assessed by examining the F-values as they return the relative variance that is explained by the term against the total variance of the data. For a random effect, however, the F-values cannot be determined. As the ‘strain-by-laboratory’-interaction constituted a random effect in the model, the F-values could not be assessed and p-values were used as a quantitative measurement for the contribution of this effect to the model (i.e., as a proxy for the F-values). The p-values are a function of the chi-square value of a Likelihood Ratio test assessing the random effect by comparing a full linear model versus a restricted linear model (i.e., without the interaction effect) and the degrees of freedom in the model. The degrees of freedom can be assumed to be the same in the analysis of both designs, since in both experimental designs the same sample size was used and the models have the same structure regarding the applied factor levels. Higher p-values of the interaction term indicate less impact of the laboratory on the strain effect and therefore, more consistency of the strain effect among laboratories and thus better reproducibility. Subsequently, the p-values of the ‘strain-by-laboratory’-interaction term of all 10 outcome measures were compared between the standardised and heterogenised design. This was done by using the Wilcoxon signed-rank test (paired, one-tailed). The Wilcoxon signed-rank test was chosen as a non-parametric test for dependent groups to provide a direct comparison of the reproducibility of the results between both designs using the same outcome measure (cf. von Kortzfleisch et al. 2020).

II.) Estimation of how often and how accurately the overall effect is predicted in the laboratories.

The performance of each experimental design to predict the overall effect size was assessed by the coverage probability (Pc) and the proportion of accurate results (Pa).

As a first step, the overall effect size of each outcome measure was estimated by conducting a random-effect meta-analysis on the data of all three laboratories independent of the experimental design. In detail, individual strain effect sizes and corresponding standard errors were calculated by applying the following linear mixed model (equation 2) to the data of each laboratory and experimental design, separately.

Written in formal terms:

*yimk* = *µ* + *ai* + *dm* + *fim* + *εimk* Equation (2)

where *i* = 1, …, *nS*, *m* = 1,…, *nB* and *k* = 1,…, *nijm*. *ai* indicates the main effect of the *i*th level of strain (treatment); *dm* represents block as a random effect *dm*~ N(0,σd2); *fim* represents strain-by-block-interaction as a random effect *fim*~ N(0,σf2) and the error term *εimk* ∼ N(0,σe2).

Written in layperson terms:

y = ‘strain’ + ‘block’ + ‘block x strain’ Equation (2)

where ‘strain’ was included as fixed effect and ‘block’ and the ‘block-by-strain’-interaction as random factors to account for the structure of the randomised block design in each laboratory (for details see Fig 2)

The random-effect meta-analysis was based on the individual strain effect sizes and standard errors of all laboratories deriving from equation 2. It was conducted using the R-package ‘metafor’ (Viechtbauer 2010, Version 2.1.0). The intercept of the following linear mixed model (equation 3) was used to return the overall effect sizes and corresponding CI95 of each outcome measure.

Written in formal terms:

*Si* =*µ* + *fi* + *εi* Equation (3)

where *i* = 1, …, *nR*. *Si* represents the estimated strain effect sizes. *fi* indicates laboratory as a random effect and the error term *εi* ∼ N(0,σe2).

Written in layperson terms:

y = ‘laboratory’ Equation (3)

where ‘laboratory’ was included as random effect to account for the variation between laboratories.

Following this step, mean strain differences and CI95 were computed based on the LMM in equation 2 using the R-package ‘lsmeans’ (Lenth & Lenth 2018, Version 2.30.0) for each laboratory and both designs, separately. Then the Pc and Pa were assessed for each experimental design. The Pc was calculated by counting how often the CI95 of the different laboratories covered the overall effect size, whereas the Pa was determined by counting how often the laboratories predicted the overall effect accurately concerning its statistical significance. For the latter, it was examined whether the CI95 of the overall pooled effect overlapped with 0 (i.e., overall not significant effect) or not (i.e. overall significant effect). In detail, two requirements had to be met for a laboratory to be counted as predicting the overall effect accurately. The CI95 of the results from one laboratory had to include the overall effect and if CI95 of the overall effect included 0, then the CI95 of the laboratory also had to include 0 (for a graphical explanation of this concept seeVoelkl et al. 2018 and von Kortzfleisch et al. 2020).

Finally, an analysis similar to the p-values of the interaction-term was conducted; the Pc and Pa ratios of all 10 outcome measures were compared between both designs using the Wilcoxon signed-rank test (paired, one-tailed).

**Components of variance**

To disentangle the different sources of variation and estimate the proportion of variance attributable to these factors, the following linear mixed model (LMM, equation 4) was applied to the full data set.

Written in formal terms:

*yijmk* = *µ* + *ai* + *bj* + cij + *dm* + *fim* + *gn* + *hin* + *kp* + *εijmr* Equation (4)

where *i* = 1, …, *nS*, *j* = 1, …, *nR*, *m* = 1,…, *nB* and *k* = 1,…, *nijm*. *ai* indicates the main effect of the *i*th level of strain (treatment); *bj* represents laboratory as a random effect where *bj*~ N(0,σb2); *cij* represents strain-by-laboratory-interaction as random effect where *cij*~ N(0,σc2); *dm* represents experimenter as a random effect nested in laboratory where *dm*~ N(0,σd2); *fim* represents strain-by-experimenter-interaction as a random effect where *fim*~ N(0,σf2); *gn* represents block as a random effect nested in laboratory and experimenter where *gn*~ N(0,σd2); *hin* represents strain-by-block-interaction as a random effect where *hin*~ N(0,σf2); *kp* represents cage as a random effect where *kp*~ N(0,σf2) and the error term *εijmr* ∼ N(0,σe2).

Written in layperson terms:

y = ‘strain’ + ‘laboratory’ + ‘experimenter’ + ‘strain x laboratory’ + ‘strain x experimenter’ + ‘block’ + ‘block x strain’ + ‘cage’ Equation (4)

where ‘strain’ was included as fixed factor and ‘laboratory’, ‘experimenter’, ‘strain-by-laboratory’-interaction, ‘strain-by-experimenter’-interaction, ‘block’, ‘block-by-strain’-interaction and ‘cage’ as random factors. The factor ‘block’ was nested within ‘experimenter’ which was nested within ‘laboratory’. For the two outcome measures Nest test 5h score and Nest test 24h score, the random factor ‘cage’ was not included in the LMM, because scores were assessed at the cage level. For the component of variance analysis, the amount of estimated variance for each component in the LMM was calculated and divided by the total variation of the outcome measure to provide the relative explained variation by each factor.

References:

Lenth, R. & Lenth, M. R. Package ‘lsmeans’. Am. Stat. 34, 216–221 (2018).

Viechtbauer, W. Conducting meta-analyses in R with the metafor package. J. Stat. Softw. 36, 1–48 (2010).

Voelkl, B., Vogt, L., Sena, E. S. & Würbel, H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLoS Biol. 16, 1–13 (2018).  
doi:10.1371/journal.pbio.2003693

von Kortzfleisch, V. T. et al. Improving reproducibility in animal research by splitting the study population into several ‘mini-experiments’. Sci. Rep. 10, 16579 (2020).