Multi-Week Digital Home Cage Monitoring Reduces Noise and Enhances Reproducibility

Michael C. Saul[1,2,*], Natalie Bratcher-Petersen[2,3,], Manuel E. Ruidiaz[2,3], Pablo Oberhauser[2,3], Vivek Philip[1,2], Susan E. Bolin[4], Brianna N. Gaskill[5], and Timothy L. Roberston[1,2,3,*].

[1]The Jackson Laboratory for Mammalian Genetics, Bar Harbor, ME, USA; [2]Digital In Vivo Alliance, Redwood City, CA, USA; [3]TLR Ventures, Redwood City, CA, USA; [4]AbbVie Inc., North Chicago, IL, USA; [5]Novartis BioMedical Research, Cambridge, MA, USA.

[*]michael.saul@jax.org, 600 Main St, Bar Harbor, ME 04609 USA.

Abstract

Reproducibility is a persistent challenge in preclinical research. We used multi-week rodent machine vision home cage monitoring at three different pharmaceutical companies to examine factors governing replication of genotype differences in activity. Interlaboratory replication of genotype effects was surprisingly high. Longer study durations reduced noise, improving replication and reducing replicable sample sizes. These findings demonstrate the potential of long-term home cage digital monitoring as a method to enhance reproducibility.

Text

Results that fail to reproduce within and replicate between laboratories pose a challenge in preclinical research. There are preventable causes of irreproducible research including over-standardization, improper or underpowered study designs, uncaptured environmental variability, experimenter influences on test subjects, and insufficient protocol specification and harmonization[1–4]. Many paradigms lack automation, are performed in non-standard environments, list methods of an hour or less in duration, and are performed during hours that disrupt the rodent sleep/wake cycle[5,6]. Moreover, even with experiments explicitly designed to overcome reproducibility issues, interlaboratory variability and unmeasured and unaccounted for experimental variation persist in challenging reproducibility[5,6].

Identifying new methods to enhance reproducibility in animal research is of great translational value. Previously suggested methods for enhancing reproducibility include reassessing species selection, utilizing systematic heterogenization that includes performing studies at different times of day, capturing cage effects, and understanding the influence of factors such as weaning time and housing conditions[7,8].

Digital home cage monitoring shows potential to improve the reproducibility of preclinical research[9]. Continuous and longer duration measurement of behavior and physiology in the animal's undisturbed home environment minimizes variability and enables assessment of the impact of human interaction. The Digital In Vivo Alliance (DIVA) is a collaboration led by The Jackson Laboratory, bringing together pharmacologists, veterinarians, machine learning experts, and data scientists to develop and validate digital measures with the goal of enhancing the translational relevance of preclinical research models[10]. The JAX Envision™ platform facilitates scalable real-time monitoring of individual animals in group housed home cage environments. DIVA has used this platform to study and improve preclinical research by enabling detailed protocol harmonization, automated continuous measurement to enhance statistical power, unbiased operator-independent assessments, and reduced human–animal interaction.

As an initiative of the DIVA Animal Health, Husbandry and Welfare focus group, we sought to understand how sources of variance in activity across three member sites contribute to measured biological effects. Unbiased digital measures collected in the home cage allow extended duration data collection measured in weeks. We posited that digital home cage phenotyping would improve inter-site replication through enhanced capture of sources of variation within and across sites. We recorded home cage activity of males and females of three mouse genotypes (inbred C57BL/6J and A/J strains as well as J:ARC, an ICR-derived Swiss outbred stock) in three different vivaria (AbbVie in Lake County, IL; Novartis in San Diego, CA; and BioMarin in Marin County, CA). Mice were housed in cages on a 14L:10D light/dark cycle with the lights-on zeitgeber at 04:00 local time for all sites and the lights-off zeitgeber at 18:00 local time for all sites. Each site performed experiments on mice shipped on three dates from the Ellsworth, Maine facility of The Jackson Laboratory. Each site used mice from the same three ship dates. Sites agreed on a study protocol and attempted to standardize as many variables as possible, including light cycle, bedding, lot/batch of rodent diet, enrichment and cage change schedule and handling practices at cage change (see: Methods).

Mice were co-housed in groups of 3 and acclimated for 3 days. After the acclimation period, mice were tagged with custom RapID® (RapID Labs, San Francisco, CA) ear tags to facilitate individual animal identification. Mice were then tracked for a further 18 days. We modeled all variables in the dataset with an ANOVA model similar to the model used by Crabbe, Wahlsten, and Dudek[5]:

$$\text{actvity} = \beta_0 + \beta_1 * \text{site} + \beta_2 * \text{ship} + \beta_3 * \text{genotype} + \beta_4 * \text{sex} + \beta_5 * (\text{genotype} \times \text{sex}) + \beta_6 * (\text{site} \times \text{genotype}) + \beta_7 * (\text{ship} \times \text{genotype}) + \varepsilon$$

Briefly, genotype, sex, and genotype-by-sex interaction were considered biological factors of interest; site, ship date, and their respective interactions with genotype to be technical and/or environmental factors inseparable from technical factors; and the residual to be measurement error.

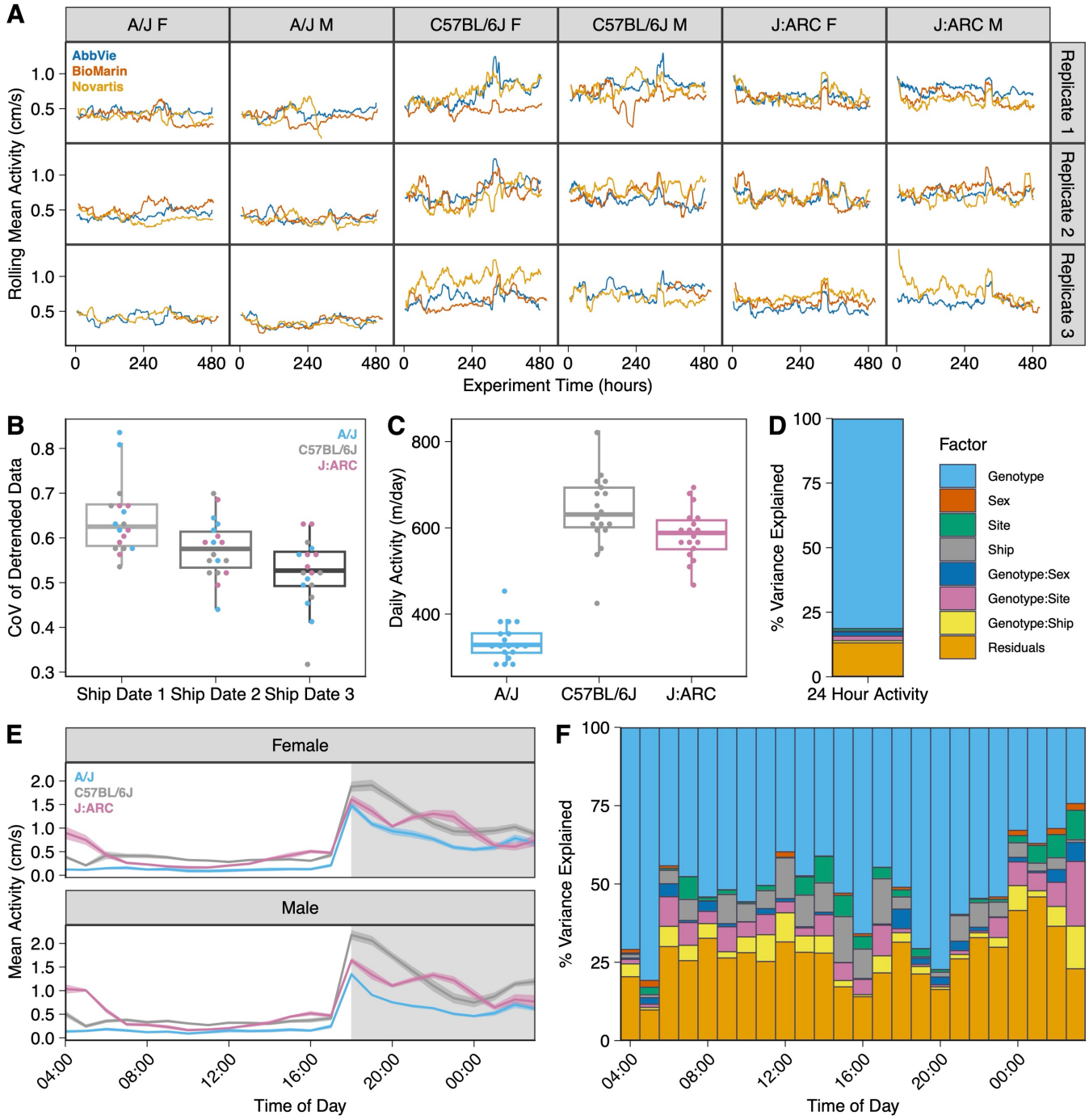

Figure 1: Effect of genotype scales with experiment duration. **A)** Hour-to-hour variation in 24-hour moving average data. **B)** Coefficient of variation of detrended data differs across ship dates. **C)** Total daily activity is strongly governed by genotype at full study duration. **D)** Genotype explains over 80% of the variance in the study. **E)** Each genotype has a unique circadian profile that is highly stereotyped. F) The strength of the genotype effect depends on time of day.

A few notable events manifested during data collection that resulted in attrition. In the first ship date, mice in cages at two sites experienced dehydration, resulting in one adverse husbandry event that precipitated early termination of the study for some cages and adjustment of hydration in all cages by the addition of Hydro-Gel® (Clear H2O, Westbrook, ME) for this and subsequent trials. During the final ship date, an error led to the loss of multiple days of video data at one site. Despite these issues, a total of 76,495 hours (8.73 years) of mouse home cage activity were

tracked across the 3 sites. The data were analyzed with cage-level activity as the primary dependent variable. We note that the experiment was designed with the cage as the experimental unit.

There was substantial variation in average activity within each cage as measured by 24-hour rolling means, a smoothing method that averages out circadian components (**Figure 1A**). Technical factors were most strongly related to the hourly variation in circadian-detrended data with ship date explaining 31.7% of the variance in the coefficient of variation for detrended hourly data within cages ($F_{2,36} = 16.47$, $p < 0.001$, **Figure 1B**). The higher variation in the earliest ship date may reflect noise introduced by the adverse husbandry events noted above.

Though variability dominated the data at the granular 1-hour timescale, a different picture emerged when the data were taken in aggregate. The genotype effect explained 81.3% of the variance in activity (**Figure 1C-D**) when 24-hour activity was averaged across all days of the experiment. All technical factors explained only 3.9% of the variance, and residuals explained 13.1% of the variance in average activity, indicating that both technical factors and error were minimal in the aggregate measures. Sex and its interaction with genotype explained the final 1.7% of the variance. Further, each site could individually identify strong genotype effects on activity (75.5%-88.7% variance explained, AbbVie: $F_{2,6} = 36.24$, $p < 0.001$; BioMarin: $F_{2,6} = 39.82$, $p < 0.001$; Novartis: $F_{2,6} = 36.11$, $p < 0.001$), which demonstrates replication of findings between these sites.

Data were further evaluated by time of day, where each hour of the day for each cage was averaged together across days and each genetic background and sex displayed its own characteristic circadian pattern (**Figure 1E**). As expected, each genotype was generally active in the dark period and inactive in the light period. In general, C57BL/6J mice showed the highest activity, J:ARC mice showed intermediate activity, and A/J mice showed the lowest measured activity, consistent with previous measurement[11]. Relative to the other two genetic backgrounds, J:ARC animals showed less attenuation of activity during the first half of the dark period. Factors influencing replication also depended upon time of day (**Figure 1F**); with the strongest genotype effects at specific moments of the day (04:00-05:00, 18:00-23:00), while residuals (00:00-01:00), and technical factors (06:00-17:00, 02:00-03:00) were stronger at different times. Thus, time of measurement influences site-to-site reproducibility of results.

To quantify how the duration of a study influences replication of the hourly results, the duration of the dataset was subsampled, generating 100 pseudo experiments for each hour from randomly sampled data with durations between 1 and 20 days (Supplementary Methods). Each pseudo experiment had approximately the same sample sizes and degrees of freedom for each model factor, allowing direct comparison of the effects on different model factors between pseudo experiment durations.

In general, shorter duration pseudo experiments were associated with greater experimental noise as measured by rescaled residual variance. As the duration of the pseudo experiments increased, rescaled variance explained by technical and biological factors also increased (**Figure 2A**). To test replication in the hourly pseudo experiments, ANOVAs were run for each site, testing for genotype effects. As pseudo experiment duration increased, significant results between sites (**Figure 2B**) were more likely to be replicated. Replication at each hour in the 20 day pseudo experiments was also examined. The ability to replicate the genotype findings depends strongly upon the time of day when tests were performed (**Figure 2C**). These results indicate that reproducibility is enhanced both by longer study duration and by using 24-hour continuous monitoring to study the times that would otherwise be unavailable when experimenters are typically present.

These effect sizes were used to perform power analysis on experiments expected to uniquely replicate the genotype findings in this experiment 80% of the time at three sites given these effect sizes ($\sqrt[3]{1-\beta}$ where $\beta = 0.2$ using one-way ANOVAs on three classes for $p = 0.05$). For a single day duration, the median number of cages needed for replicable results was n=39 with a range between n=8 at 04:00 to n=75 at 13:00. Sample sizes for single day durations appeared to be relatively high between 10:00 and 15:00, (**Figure 2D**), a period when people are present in the animal room and when data collection is often performed. For 20 days' duration, the median sample size for replicable results reduced to n=8 and ranged from n=3 at 05:00 to n=23 and 03:00 (**Figure 2D**). The results of the power analysis indicate that short duration testing for 1 hour or less during the time when experiments are

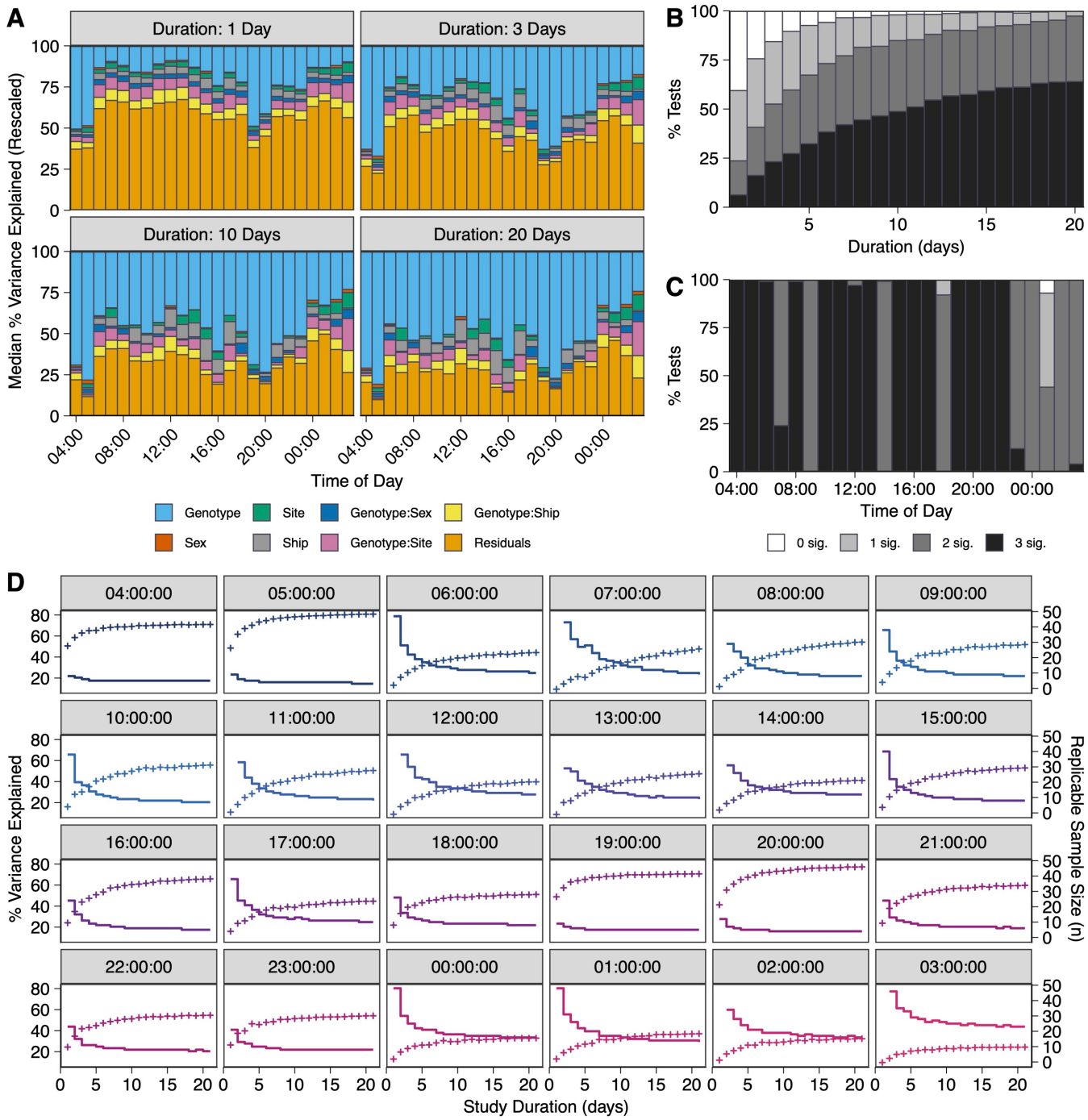

Figure 2: Results of hourly pseudo experiments. **A)** Residual variance decreases with longer pseudo experiment duration. **B)** Longer pseudo experiments increase likelihood of multiple site replication of genotype findings. Darker colors indicate replication of statistically significant hypothesis test results at more sites. Black indicates replication at all three sites. **C)** Multi-site replication with 20-day pseudo experiments depends upon time of day. **D)** Percent variance explained (crosses) increases with pseudo experiment duration. Sample size needed for well-powered replicable experiments (solid stair-stepped line) strongly decreases as duration increases.

typically performed requires very high sample sizes to achieve replicable results, but that sample sizes may be reduced by looking at different circadian times and by averaging over many days.

Though substantial noise and technical variation at short durations were discovered, long-duration monitoring allowed us to overcome these issues and increase inter-laboratory replication. These results demonstrate that multi-week designs with many data points per biological replicate averaged together strongly enhance reproducibility. Further, reproducibility appears to be improved with digital home cage measures, which facilitate unbiased measurement of behavioral and physiological endpoints for long durations that include hours when animals are typically awake, and people are not present in the animal room. We note limitations of these findings, specifically for identifying whether increased variability in the light period resulted from circadian influences or from the presence of humans in the room at these times. Future work on reproducibility would benefit from examining such additional sources of variability that we were unable to assess here.

References

1. Nigri, M., Åhlgren, J., Wolfer, D. P. & Voikar, V. Role of Environment and Experimenter in Reproducibility of Behavioral Studies With Laboratory Mice. *Front. Behav. Neurosci.* **16**, (2022).
2. Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLoS Med.* **2**, e124 (2005).
3. 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).
4. Karp, N. A. *et al.* A multi-batch design to deliver robust estimates of efficacy and reduce animal use – a syngeneic tumour case study. *Sci Rep* **10**, 6178 (2020).
5. Crabbe, J. C., Wahlsten, D. & Dudek, B. C. Genetics of mouse behavior: interactions with laboratory environment. *Science* **284**, 1670–1672 (1999).
6. Von Kortzfleisch, V. T. *et al.* Do multiple experimenters improve the reproducibility of animal studies? *PLoS Biol* **20**, e3001564 (2022).
7. Bodden, C. *et al.* Heterogenising study samples across testing time improves reproducibility of behavioural data. *Sci Rep* **9**, 8247 (2019).
8. Bailoo, J. D. *et al.* Effects of weaning age and housing conditions on phenotypic differences in mice. *Sci Rep* **10**, 11684 (2020).
9. Pernold, K. *et al.* Towards large scale automated cage monitoring - Diurnal rhythm and impact of interventions on in-cage activity of C57BL/6J mice recorded 24/7 with a non-disrupting capacitive-based technique. *PLoS One* **14**, e0211063 (2019).
10. Digital In Vivo Alliance. *DIVA* https://diva.bio/home/.
11. Logan, R. W. *et al.* High-precision genetic mapping of behavioral traits in the diversity outbred mouse population: Genetic mapping of behavioral traits in the outbred mouse. *Genes, Brain and Behavior* **12**, 424–437 (2013).
12. Robertson, T. L. *et al.* An integrated and scalable rodent cage system enabling continuous computer vision-based behavioral analysis and AI-enhanced digital biomarker development. 2024.12.18.629281 Preprint at https://doi.org/10.1101/2024.12.18.629281 (2024).

Acknowledgements

This study was a collaborative effort by members of the Digital In Vivo Alliance (DIVA). We extend our sincere thanks to Kristin Evans and Brian Baridon for leading the study and conducting the work at the BioMarin site, and to Christina Boykin and Joseph Brewster for leading and conducting the work at the Novartis site. We are also grateful to the AbbVie research training services team for handling anesthesia and ear tagging, and to the husbandry team for their precise execution of health monitoring and cage changes. Special thanks to Brian Berridge (DIVA Chair, B2 Pathology Solutions) for his thoughtful feedback on an earlier manuscript draft and for raising insightful questions that guided meaningful revisions. We also acknowledge the Jackson Laboratory for providing the mice and the beta version of the technology used in data collection.

Declaration of Interests

Timothy L. Robertson is the founder, CEO, and owner of TLR Ventures, which developed the technology utilized in this work. Natalie Bratcher-Petersen, Manny Ruidiaz, and Pablo Oberhauser, are employees of TLR Ventures. Natalie Bratcher-Petersen is also the manager of the Digital In Vivo Alliance (DIVA). Michael Saul is an employee of The Jackson Laboratory, which commercialized this technology. Susan E. Bolin is an employee of AbbVie, Inc., and AbbVie is a member of DIVA. Brianna N. Gaskill is an employee of Novartis, and Novartis is a member of DIVA.

Author Contributions

NB-P, SB, BNG, and MCS jointly designed the study, MR, JPO, and TLR developed the machine vision tools used to identify mice and track activity. MCS analyzed the data and wrote statistical analysis code. VMP provided critical guidance on statistical analysis. MCS drafted the manuscript. MCS, NB-P, MR, JPO, SB, AM, BNG, and TLR reviewed, revised, and edited the manuscript. NB-P administered project logistics. TLR provided funding through DIVA.