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], Juan Pablo Oberhauser[2,3], Vivek M. 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.

* **Direct correspondence to:** Michael C. Saul, email**:** michael.saul@jax.org

Author Contributions: MCS, NB-P, SEB BNG,, and TLR designed research. MCS, NB-P, SEB, BNG, and TLR performed research. MCS, MER, JPO, and TLR contributed new analytic tools. MCS and VMP analyzed data. MCS, NB-P, SEB, BNG, and TLR wrote the paper.

Competing Interest Statement: TLR is the CEO of TLR Ventures, which developed the technology utilized in this work. NB-P, MER, and JPO, are employees of TLR Ventures. NB-P is the manager of the Digital In Vivo Alliance (DIVA). MCS is an employee of The Jackson Laboratory, which commercialized this technology. SEB and BNG are employees of DIVA members. MER, JPO, and TLR are co-inventors on a patent on the underlying technology utilized in this manuscript.

Main Text

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.

Introduction

Results that fail to reproduce pose a challenge in preclinical research. Over-standardization, improper or underpowered study designs, uncaptured environmental variability, experimenter influences on subjects, and insufficient protocol specification and harmonization can cause irreproducible research (Karp et al., 2020; Nigri et al., 2022; Richter et al., 2009). Paradigms that are not automated, performed in irregular environments, use short-duration methods, and occur during hours mice would typically be asleep further complicate reproducibility (Crabbe et al., 1999; Von Kortzfleisch et al., 2022). Further, unmeasured environmental factors and interlaboratory variation complicate reproducibility in ideal experiments (Crabbe et al., 1999; Von Kortzfleisch et al., 2022).

Enhancing reproducibility in preclinical research has great translational value. Suggested methods for enhancing reproducibility include reassessing species selection, utilizing systematic heterogenization, performing studies at different times of day, capturing cage effects, and controlling for factors like weaning time and housing (Bailoo et al., 2020; Bodden et al., 2019).

One potential method to improve preclinical research reproducibility is digital home cage monitoring (Pernold et al., 2019). Continuous long-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), a collaboration led by The Jackson Laboratory, aims to develop and validate digital measures and enhance translational and preclinical research. The JAX Envision™ platform facilitates scalable real-time monitoring of group-housed individuals in home cage environments. DIVA has used Envision™ for detailed protocol harmonization, automated continuous measurement, unbiased operator-independent assessments, and to reduce human–animal interaction.

We sought to understand how variation across three member sites contribute to measured biological effects in home cage activity. 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.

Brief Methods

We recorded home cage activity of both sexes of three mouse genotypes (C57BL/6J, A/J, and J:ARC) in three different vivaria (AbbVie in Lake County, IL; Novartis in San Diego, CA; and BioMarin in Marin County, CA). Mice were housed in a 14L:10D light/dark cycle with lights-on at 04:00 local time and lights-off at 18:00 local time for all sites. All mice shipped from the Ellsworth, Maine facility of The Jackson Laboratory on three ship dates. Sites agreed on a study protocol and standardized as many variables as possible (light cycle, bedding, lot/batch of rodent diet, enrichment, cage change schedule, and handling practices; see: **Supplementary Methods**).

Mice were housed in groups of three, acclimated for three days, then tagged with custom RapID® tags (RapID Labs, San Francisco, CA) for individual animal identification. Mice were tracked for a further 18 days. We utilized an ANOVA model from previously published work (Crabbe et al., 1999):

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

Genotype, sex, and their interaction were considered biological factors of interest; site, ship date, and their interactions with genotype were considered technical/environmental factors.

Results and Discussion

We tracked a total of 76,495 hours (8.73 years) of mouse home cage activity across the 3 sites. During the first ship date, mice two sites experienced dehydration, resulting in early study termination for one cages and addition of Hydro-Gel® (Clear H2O, Westbrook, ME) for this and subsequent trials. During the final ship date, multiple days of video data were lost due to an error.

The experiment was designed with the cage as the experimental unit and data were analyzed with cage-level activity as the primary dependent variable. There was substantial variation in average activity within each cage as measured by 24-hour rolling means (**Figure 1A**). Ship date explained 31.7% of the variance in the hourly coefficient of variation for detrended data ($F_{2,36}$ = 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.

When the data were taken in aggregate, genotype explained 81.3% of the variance in average 24-hour activity (**Figure 1C-D**). Technical factors explained 3.9% of the variance and residuals explained 13.1% of the variance in average activity, indicating that both technical factors and noise were minimal in the aggregate measures. Sex and its interaction with genotype explained the final 1.7% of the variance. Further, each site independently identified strong genotype effects on 24-hour 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, $p < 0.001$), demonstrating replication between these sites.

Data were further averaged by time of day. Each genetic background and sex displayed its own characteristic circadian pattern (**Figure 1E**). Each genotype was active in the dark period and inactive in the light period. Consistent with previous measurement (Logan et al., 2013), C57BL/6J generally showed the highest activity, J:ARC was intermediate, and A/J was lowest. J:ARC activity remained relatively stable across the dark period. Genotype effect strength depended upon time of day (**Figure 1F**); genotype effects dominated at specific times (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 other times. Thus, time of measurement influences site-to-site reproducibility of results.

To simulate how the duration of a study influences replication, each hour was randomly subsampled 100 times to between 1 and 20 days (**Supplementary Methods**). Each subsample yielded approximately the same sample size, allowing direct comparison between experiment durations.

Shorter duration was generally associated with greater experimental noise. As subsample increased, rescaled variance explained by technical and biological factors also increased (**Figure 2A**). ANOVAs for each site demonstrate that replication as measured by simultaneous

significant results between sites increased for subsamples with longer durations (**Figure 2B**). In 20 day subsamples, genotype effect replication varied by 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 capture replicable times when experimenters would typically be absent.

Subsample effect sizes were used to perform power analysis for sample sizes needed to replicate across three sites ($\sqrt[3]{1-\beta}$ 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 relatively high between 10:00 and 15:00, (**Figure 2D**), when workers are often present and when data is often collected. 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 typically performed requires very high sample sizes to achieve replicable results, but that sample sizes may be reduced both by testing at different circadian times and by averaging over many days.

Long-duration monitoring allowed us to overcome substantial day-to-day noise and increase inter-laboratory replication. Multi-week designs with many data points per biological replicate averaged together enhance reproducibility. Further, reproducibility appears to be improved through unbiased measurement of behavioral and physiological endpoints during hours when animals are typically awake and humans are not present in the animal room. We note that these findings suggest that increased variability in the light period resulted from circadian influences or from the presence of humans in the room at these times, but that the present dataset cannot adjudicate this hypothesis. Future work on reproducibility would benefit from examining such additional sources of variability that we were unable to assess here.

Acknowledgments

We than Kristin Evans and Brian Baridon from BioMarin, Christina Boykin and Joseph Brewster from Novartis, the AbbVie research training services, and the three site animal husbandry teams. We thank Brian Berridge of B2 Pathology Solutions and DIVA chair for his thoughtful feedback. We acknowledge the Jackson Laboratory for providing the mice and the beta version of the technology used in data collection and DIVA for funding the work.

References

Bailoo JD, Voelkl B, Varholick J, Novak J, Murphy E, Rosso M, Palme R, Würbel H. 2020. Effects of weaning age and housing conditions on phenotypic differences in mice. Scientific Reports **10**:11684. DOI: https://doi.org/10.1038/s41598-020-68549-3

Bodden C, von Kortzfleisch VT, Karwinkel F, Kaiser S, Sachser N, Richter SH. 2019. Heterogenising study samples across testing time improves reproducibility of behavioural data. Scientific Reports **9**:8247. DOI: https://doi.org/10.1038/s41598-019-44705-2

Crabbe JC, Wahlsten D, Dudek BC. 1999. Genetics of mouse behavior: interactions with laboratory environment. Science **284**:1670–1672.

Karp NA, Wilson Z, Stalker E, Mooney L, Lazic SE, Zhang B, Hardaker E. 2020. A multi-batch design to deliver robust estimates of efficacy and reduce animal use – a syngeneic tumour case study. Scientific Reports **10**:6178. DOI: https://doi.org/10.1038/s41598-020-62509-7

Logan RW, Robledo RF, Recla JM, Philip VM, Bubier JA, Jay JJ, Harwood C, Wilcox T, Gatti DM, Bult CJ, Churchill GA, Chesler EJ. 2013. 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. DOI: https://doi.org/10.1111/gbb.12029

Nigri M, Åhlgren J, Wolfer DP, Voikar V. 2022. Role of Environment and Experimenter in Reproducibility of Behavioral Studies With Laboratory Mice. Frontiers in Behavioral Neuroscience **16**. DOI: https://doi.org/10.3389/fnbeh.2022.835444

Pernold K, Iannello F, Low BE, Rigamonti M, Rosati G, Scavizzi F, Wang J, Raspa M, Wiles MV, Ulfhake B. 2019. 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. DOI: https://doi.org/10.1371/journal.pone.0211063, PMID: 30716111

Richter SH, Garner JP, Würbel H. 2009. Environmental standardization: cure or cause of poor reproducibility in animal experiments? Nature Methods **6**:257–261. DOI: https://doi.org/10.1038/nmeth.1312

Von Kortzfleisch VT, Ambrée O, Karp NA, Meyer N, Novak J, Palme R, Rosso M, Touma C, Würbel H, Kaiser S, Sachser N, Richter SH. 2022. Do multiple experimenters improve the reproducibility of animal studies? PLOS Biology **20**:e3001564. DOI: https://doi.org/10.1371/journal.pbio.3001564

Figures

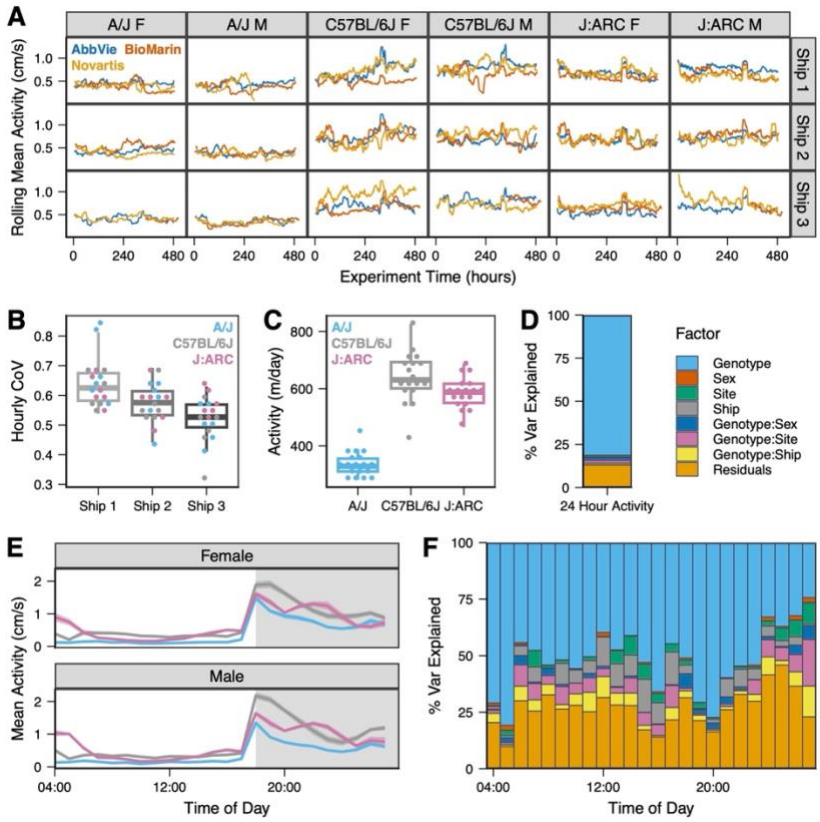

Figure 1. Genotype effect scales with duration. **A)** Variation in 24-hour moving average data. **B)** Detrended CoV differs across ship dates. **C)** Genotype strongly governs daily activity at full study duration. **D)** Genotype explains over 80% of the variance in daily activity. **E)** Genotypes have characteristic highly stereotyped circadian patterns. **F)** Genotype effect strength depends on time of day.

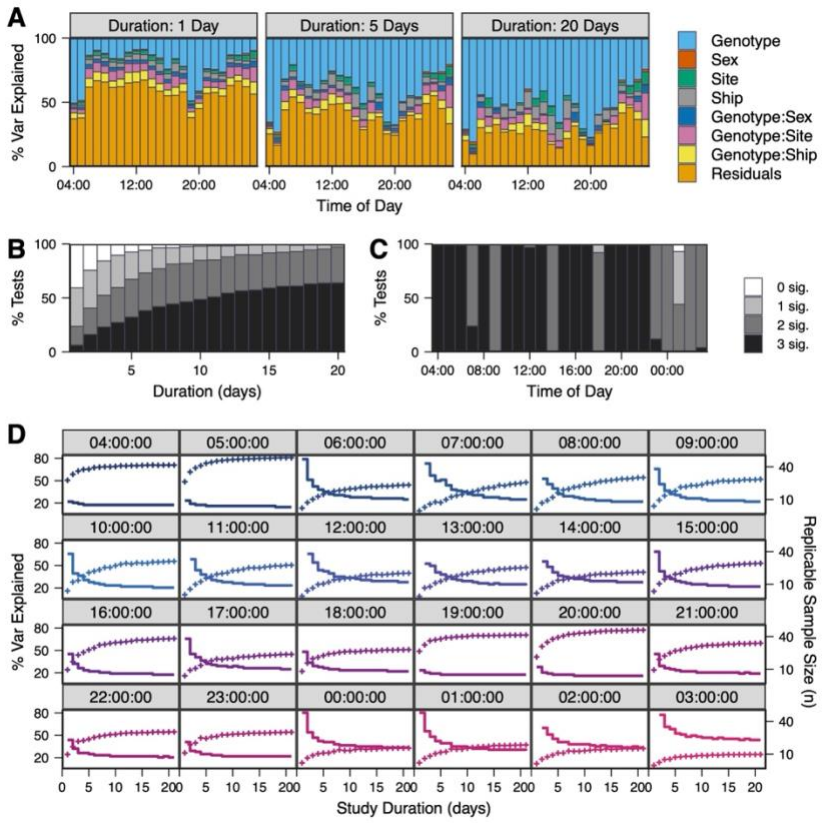

Figure 2. Results of hourly subsamples. **A)** Residual variance decreases with longer subsample duration. **B)** Longer subsamples increase likelihood of genotype effect replication. Black indicates replication at all three sites. **C)** Multi-site replication with 20-day subsamples depends upon time of day. **D)** Percent variance explained (crosses) increases with subsample duration. Sample size needed for well-powered replicable experiments (solid stair-stepped line) decreases as duration increases.