What's *Always* Wrong with my Mouse?

N. Kafkafi[1], T. Lahav[1] and Y. Benjamini[1]

Department of Statistics and Operations Research, The Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel. nkafkafi@gmail.com

In recent years there has been a growing voice of concern that a considerable percentage of published scientific discoveries fail to replicate in subsequent studies. The issue is especially relevant to preclinical studies and animal models, and has recently led to reconsideration of policies by NIH [1], as well as by some scientific journals including *Science* [2] and *Nature* [3]. Behavioral phenotyping results especially seem to be sensitive, and studies comparing inbred strains and genetically-engineered mutants across laboratories demonstrated some disturbing discrepancies [4]. These discrepancies are all the more worrying in light of the current community effort, coordinated by the International Mouse Phenotyping Consortium (IMPC), to phenotype thousands of mouse mutant lines across many laboratories during the next several years, and make the results available in public databases [5], as part of a long-term goal to functionally annotate all mammalian protein-coding genes.

While it is obvious that *something* should be done about the problem, it is less clear *what*. The new policies mostly advocate general methodological procedures and considerations, such as reporting detailed methods, pre-registering studies before the experiment and committing to sample sizes. However, they usually do not propose new statistical criteria and tools specifically designed to address the issue. Unfortunately, the intuitive notion of replicability as a central dogma of modern science has never been explicitly formulated. As recently noted by a statistician in a debate regarding replicating mouse phenotyping results: "The concept of reproducibility has not been well developed in the statistical literature, and so it is no wonder that debates like this have arisen" [6].

When estimating the difference between standardized mouse genotypes (e.g., an animal model knockout and its wild-type control) across several laboratories, the traditional criterion mostly used for a lack of phenotyping replicability is still the statistical significance of Genotype × Laboratory interaction ($G\times L$), although it is in fact misleading: it punishes high-quality behavioral measures in which the lower noise in measuring the individual animal effect ("within group") increases the power to discover all other effects: Genotype, Laboratory and $G\times L$ [7]. The other side of this problem is that low-quality, noisy measures might mistakenly appear replicable, if they fail to discover $G\times L$ while just discovering some strong genotype differences. We therefore argue that the more appropriate statistical model is mixed model ANOVA, in which the laboratories are regarded as a random variable ("random lab model" or RLM, as opposed to the traditional "fixed lab model" or FLM). RLM considers the laboratories in the study as *a sample*, representing the population of all potential phenotyping labs out there. It therefore adds the $G\times L$ "noise" to the individual animal noise as the yardstick against which genotype differences are judged. In practical terms, adopting RLM means raising the benchmark for showing a significant genotype effect, thus trading some statistical power for ensuring replicability [7].

In order to further examine the relevance of the FLM and RLM for replicability across laboratories we analyzed behavioral results from several mouse phenotyping studies, each conducted across several laboratories. Using these data we demonstrate that the commonlly-used FLM analysis frequently generates inconsistent conclusions that do not correspond with the intuitive concept of replicability. A typical example is seen in Figure 1, which shows a comparison between two genotypes, C57BL/6 and DBA/2, in the total path moved in the Elevated Zero Maze across 6 laboratories. FLM analysis indicated that C57BL/6 was significantly more active than DBA/2 across all laboratories ($p<0.05$), while RLM did not discover significant differences ($p=0.47$). Note that in 2 out of the 6 labs the DBA/2 mean was actually higher. Even worse, within one of these laboratories the DBA/2 mean was *significantly* higher, as indicated by the commonly-used t-test within this lab, although in 3 other labs it was significantly lower.

This kind of inconsistency in FLM analysis is not rare: in this dataset it was found in 30% of the measures in which genotype difference across all laboratories was significant in the FLM but not in the RLM. In contrast it

Proceedings of Measuring Behavior 2014, (Wageningen, The Netherlands, August 27-29, 2014).
Editors: A.J. Spink, L.W.S. Loijens, M. Woloszynowska-Fraser & L.P.J.J. Noldus. www.measuringbehavior.org

was found in none of the measures in which RLM too indicated a significant effect across all laboratories. The same comparison in another dataset, the "heterogenized" dataset from Richter et al. [8], revealed an even worse result for the FLM: 40% vs. none.

In conclusion, our examination of the data reveals that the commonly-used statistical model, in which the laboratory is treated as a fixed variable, should not be used for estimating replicability of phenotyping results across laboratories. Instead we recommend using the significance of the genotype difference in a model that treats the laboratory as a random variable.

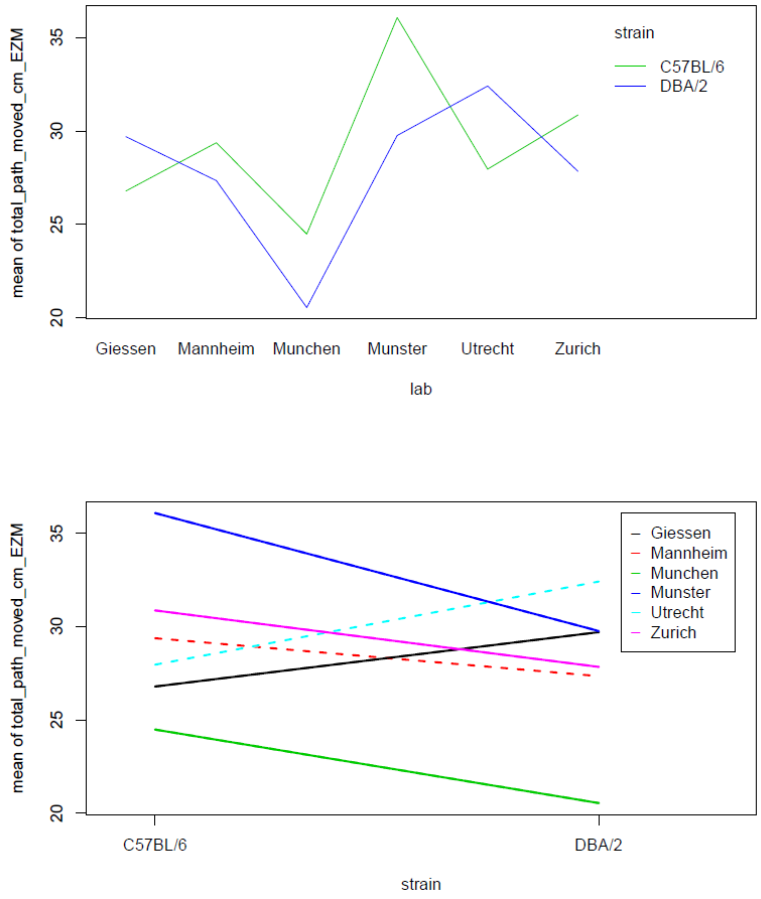

Figure 1: Differences between mouse genotypes C57BL/6 and DBA/2 across 6 laboratories, in the path moved in the Elevated Zero Maze, out of the Richter et al 2011 [8] "standardized" dataset. The same group means are are according to genotype (top) and to laboratory (bottom), and they are connected by lines in order to visualize Genotype × Laboratory interaction, seen as different and occationly even opposite slopes. In the bottom graph, continuous lines indicate genotype differences that were significant (p<0.05) using two-tailed t-test within the corresponding laboratory, and dashed lines denote non-significant differences.

Acknowledgements: This work is supported by a European Research Council grant PSARPS. The data from Richter et al. [8] and Wolfer et al. [9] were generously contributed to us by Prof. Würbel.

References:

1. Collins, F. S. , Tabak, L.A. (2014). Policy: NIH plans to enhance reproducibility. *Nature* **505**:612–613. doi:10.1038/505612a.

2. McNutt, M. (2014). Reproducibility: *Science* Editorial. *Science* **343**, 229. doi: 10.1126/science.1250475.

3. Landis, S. C. et al. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* **490**:187–191. doi:10.1038/nature11556.

4. Crabbe JC, Wahlsten D, Dudek BC (1999) Genetics of mouse behavior: interactions with laboratory environment. *Science* **284**:1670-2.

5. Koscielny,G. et al., (2014) The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Research.* **42** (D1): D802-D809. doi: 10.1093/nar/gkt977.

6. Wolfinger, E. D. (2013). Reanalysis of Richter et al. (2010) on reproducibility. *Nature Methods* **10**:373–374 doi:10.1038/nmeth.2438.

7. Kafkafi, N., Benjamini, Y., Sakov, A., Elmer, G.I., Golani, I. (2005). Genotype-environment interactions in mouse behavior: a way out of the problem. *Proceedings of the National Academy of Sciences U S A.* **102**:4619-4624.

8. Richter et al. (2011) Effect of population heterogenization on the reproducibility of mouse behavior: a multi-laboratory study. *PLoS One* **6**:e16461. doi: 10.1371/journal.pone.0016461.

9. Wolfer, D. P., Litvin, O., Morf, S., Nitsch, R. M., Lipp, H. P., Würbel, H. (2004). Laboratory animal welfare: cage enrichment and mouse behaviour. *Nature* **432**:821-2.