

FORMAL COMMENT

Realism and robustness require increased sample size when studying both sexes

Szymon M. Drobniak^{1,2*}, Malgorzata Lagisz^{1,3}, Yefeng Yang¹, Shinichi Nakagawa^{1,3*}

1 Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, Australia, **2** Institute of Environmental Sciences, Jagiellonian University, Kraków, Poland, **3** Theoretical Sciences Visiting Program, Okinawa Institute of Science and Technology Graduate University, Onna, Japan

* szymek.drobniak@gmail.com (SMD); s.nakagawa@unsw.edu.au (SN)



OPEN ACCESS

Citation: Drobniak SM, Lagisz M, Yang Y, Nakagawa S (2024) Realism and robustness require increased sample size when studying both sexes. *PLoS Biol* 22(4): e3002456. <https://doi.org/10.1371/journal.pbio.3002456>

Received: August 4, 2023

Accepted: December 1, 2023

Published: April 11, 2024

Copyright: © 2024 Drobniak et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Australian Research Council Discovery Project Grant (DP210100812), awarded to SN & ML, Australian Research Council Discovery Early Career Award (DE180100202), awarded to SMD, and Polish National Science Centre OPUS grant (UMO-2020/39/B/NZ8/01274), awarded to SMD. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Female subjects have been historically excluded from biomedicine and other related areas of study [1]. Such exclusion has disadvantaged females and prevented a fuller understanding of biology. Therefore, in 2016, the National Institute of Health (NIH) mandated that all NIH-funded animal and human studies consider sex as a biological variable (e.g., [2]). Yet, sex as a biological variable has not been welcomed with open arms, most likely because many researchers believe they need to increase the overall sample size with 2 sexes compared to using only 1 sex (e.g., [2]). Recently, Philips and colleagues published a *PLoS Biology* article titled “Statistical simulations show that scientists need not increase overall sample size by default when including both sexes in *in vivo* studies” [3]. As indicated in their title, the authors have concluded and recommended no increase in sample size with both sexes, which was based on a set of simulations exploring a simple but—as they claim—likely biological scenario. Their conclusion is great news for researchers who feared coping with increased experiment sizes and costs.

However, Philips and colleagues have assumed homoscedasticity between the 2 sexes, meaning variances or standard deviations of the sexes are the same throughout their simulations. Here, we first explain why such an assumption is biologically unrealistic and why heteroscedasticity between 2 sexes should be the norm rather than the exception by pointing out a wealth of empirical evidence and evolutionary arguments. We then show the results from a simulation study expanding Philips and colleagues’ work by incorporating heteroscedasticity. Our results clearly indicate that we need to increase the overall sample size to have robust statistical inference. Further, we provide statistical recommendations to deal with heteroscedasticity. We also briefly touch on what funding agencies and ethics committees can do, given our results.

Why do 2 sexes have different variances?

There are 2 major reasons why males and females have different variances in their traits and responses. The one is due to Taylor’s law, where an increase in variance accompanies an increase in a mean. The law was first used to describe organismal aggregation patterns, but this mean–variance relationship seems to be ubiquitous [4]. Often correlations are over 0.9 between mean and variance (standard deviation) on the logarithm scale, as shown in an example data from a meta-analysis on rodent diet manipulations [5] (Fig 1A–1C). Taylor’s law means that when there are sex differences in mean, there are also unavoidable differences in variances (i.e., heteroscedasticity). In their simulation, Philips and colleagues showed that a sex difference in treatment effects would usually increase statistical power. Yet, heteroscedasticity

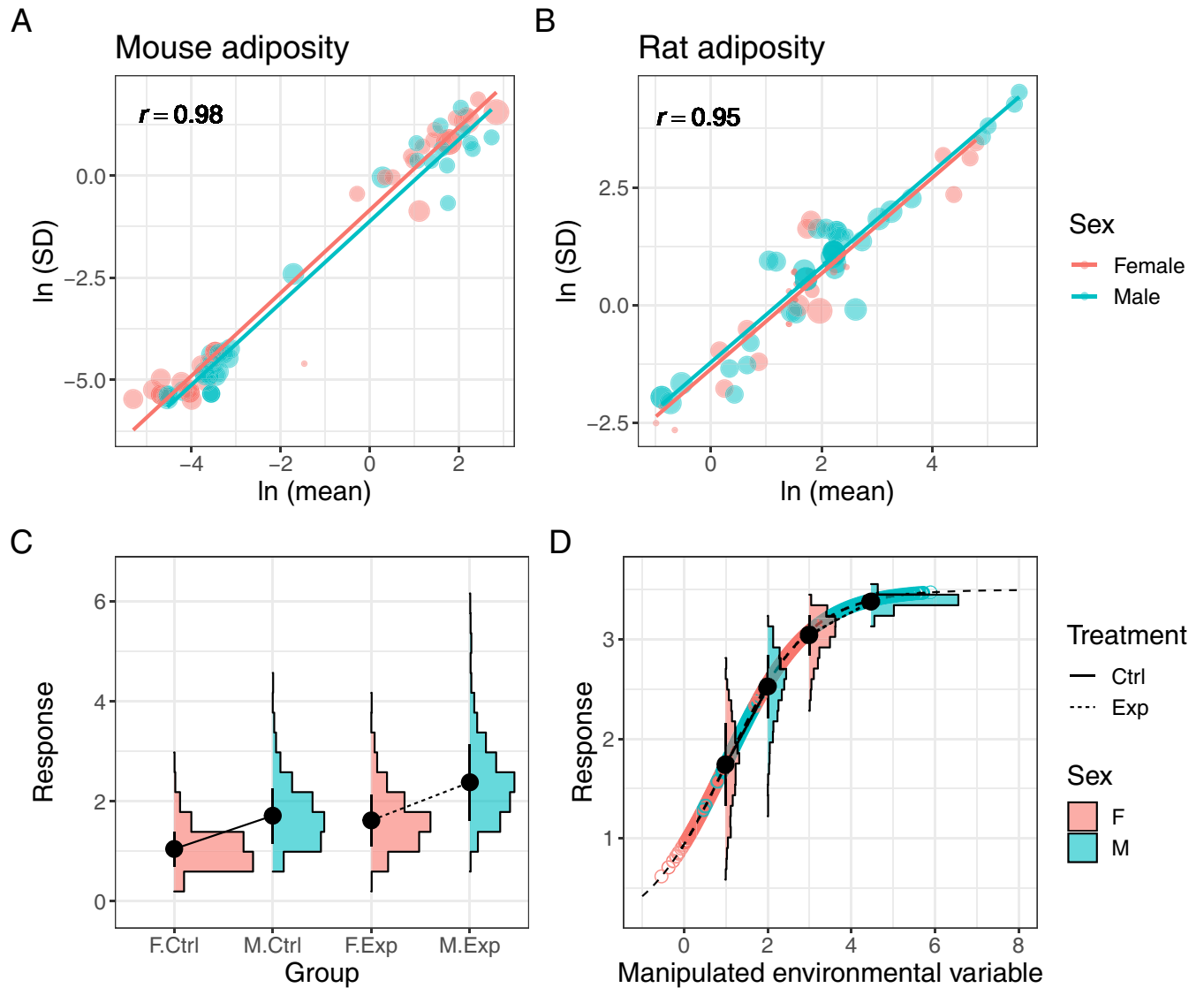


Fig 1. (A, B) An example of a strong mean and variance relationship or Taylor’s law. The offspring adiposity data is taken from a meta-analysis of 12 studies on the transgenerational effects of obesogenic diets in mice and rats [5]. Data for female and male offspring is plotted separately, correlations are calculated for both sexes, point size represents sample sizes. **(C, D) Heteroscedasticity resulting from the properties of data-generating processes.** Heteroscedastic variances may be associated with varying means if underlying distribution exhibits a mean–variance relationship (C, here a log-normal example), or if ceiling/floor effects result from nonlinear functions mapping treatment/sex effects into the phenotypic space (D, here assuming logistic mapping). See Supplementary Information for details (the data underlying this figure can be found in <https://doi.org/10.5281/zenodo.10205440>).

<https://doi.org/10.1371/journal.pbio.3002456.g001>

reduces power (see the next section). Given the empirically observed widespread mean–variance relationship, biologists often use CV (coefficient of variation; a mean standardised standard deviation) to compare variability among traits and responses. The other reason is an evolutionary inevitability, where 2 sexes have been subject to different natural and sexual selection forces. Indeed, the evolutionary and biomedical literature comparing CVs has found clear and widespread differences between sexes [6,7]. Of relevance, we have demonstrated this very point in mice [8,9]. Therefore, even if there are no sex differences in mean, we should expect heteroscedasticity between the 2 sexes.

Building upon the simulation study

As mentioned above, we conducted a simulation study in which we added, to Philips and colleagues’ simulation, 2 more scenarios with “small” and “large” heteroscedasticity, where one sex had approximately 44% and 73% larger standard deviation than the other (50% and 100% increase in sex-specific variance, respectively). As shown in Fig 2A and 2B, statistical power is lower, sometimes substantially so, when heteroscedasticity exists. This pattern is consistently true regardless of whether there is an interaction between sex and treatment. Furthermore, we investigated to see how much larger sample sizes are required when 2 sexes are heteroscedastic.

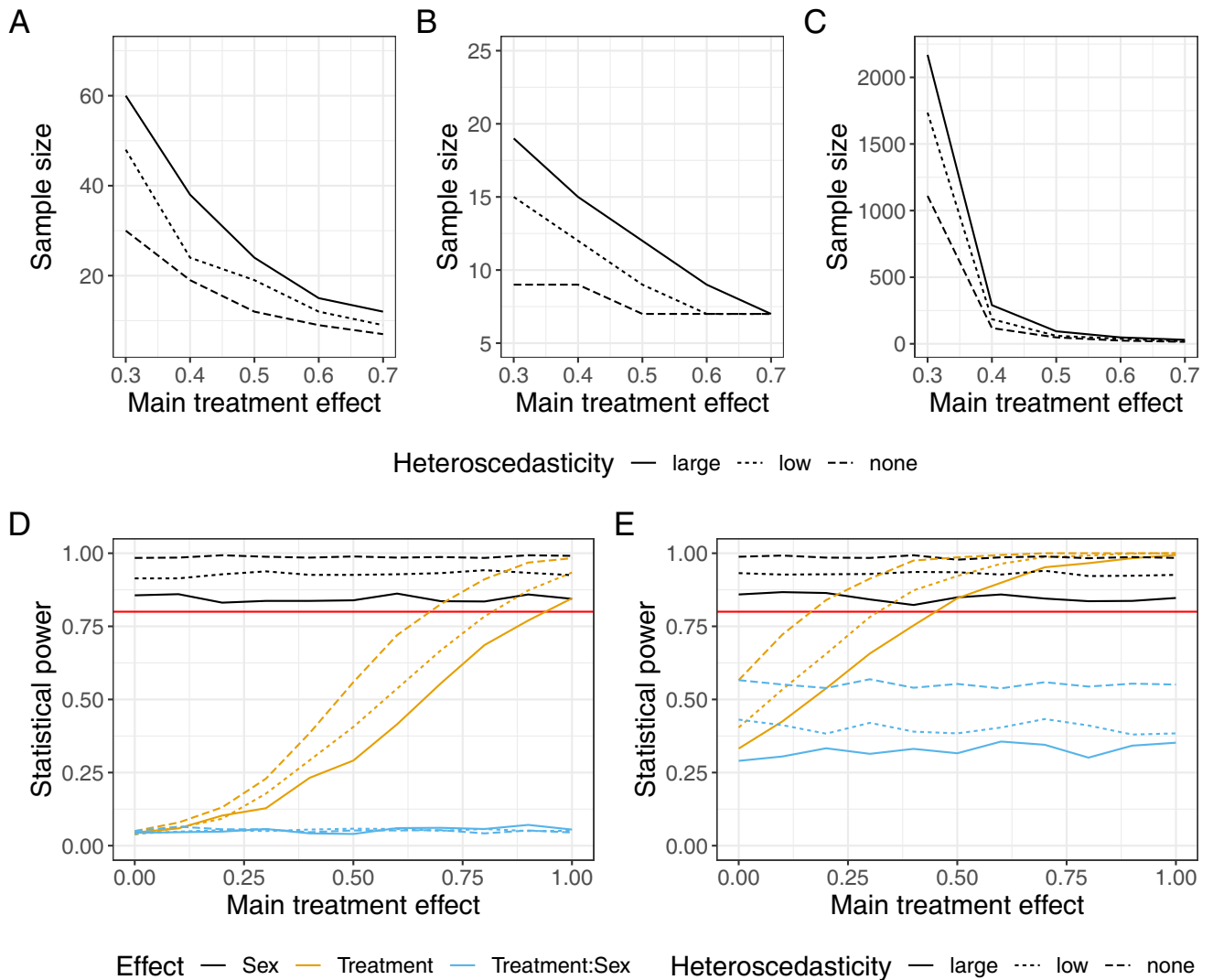


Fig 2. Simulation results. Top row: sample sizes required to achieve power of at least 80% under 3 heteroscedasticity scenarios (line type coding) and for varying magnitudes of treatment effect (x axis, limited to moderate-size effects); (A)—no sex-by-treatment interaction, (B)—interaction present, stronger treatment effect in more variables sex, (C)—same, but stronger treatment effect in less variable sex. Bottom row: power to estimate model effects (colour-coded) under 3 heteroscedasticity scenarios (line type) without (D) and with (E) sex-by-treatment interaction, with varying treatment effect size magnitude (x axis); the simulated interaction modified the magnitude of sex effect between treatment groups, but not its direction. Solid red line—power threshold of 80% (the data underlying this figure can be found in <https://doi.org/10.5281/zenodo.10205440>).

<https://doi.org/10.1371/journal.pbio.3002456.g002>

To achieve the nominal 80% power, the large heteroscedasticity scenario required twice as many sample sizes for a range of effect sizes, regardless of the interaction (Fig 2C–2E). This is especially true when the effect size (d or standardised mean difference) is less than 0.5, which would be common in non-pharmacological/toxicological studies. Importantly, the increased sample size requirements are not alleviated by using methods designed to account for heteroscedastic variance (for explanations of such methods and additional simulations using the methods, see the next section).

We note that our scenarios may represent the more extreme heteroscedasticity cases, but that they are not by any means rare or exceptional. Nevertheless, to test the robustness of our finding, we added an extra condition, constraining average variance between the sexes to remain approximately constant, and we found that such condition also leads to often larger sample size requirements in sex-heteroscedastic scenarios. Therefore, there is a clear need for larger sample sizes in a study measuring heteroscedastic traits and responses, especially when 2 sexes have different average responses. Taken together, we arrive at 2 conclusions under biological realistic scenarios (at least under the scenarios we simulated) or, i.e., with the presence of heteroscedasticity: (1) if one assumes homoscedasticity like Philips and colleagues, one's study is likely to be underpowered; and (2) if one does not model or deal with heteroscedasticity in analysis, it may lead to inflated Type I error rates exceeding 5%. All relevant R code and simulation results are found in the supplementary code document available via the webpage (see Supplementary Information, webpage: https://szymekdr.github.io/230713_power_sex_sim/).

Empirical and analytical recommendations

Therefore, we suggest researchers consider increasing sample size when including both sexes in their studies, unlike what Philips and colleagues recommended. In earlier work, we recommend that a study consider increasing the sample size for the sex with higher variability, as this is more efficient in gaining more power than increasing sample sizes in a balanced manner [8]. Also, there are 2 ways to deal with data with heteroscedasticity [10]. First, we can use heteroscedasticity-consistent standard error (variance) estimators, often known as “sandwich estimators,” with which standard errors are calculated correctly, usually increasing standard error or confidence intervals, maintaining the nominal Type I error rate (note that in many statistical software and packages, sandwich estimators are available and there are many types of sandwich estimators, e.g., [11]). Second, we can quantify homoscedasticity explicitly, using a generalized least squares model (for more details, see [10]; alternatively, one can use location-scale models; for example, [12]), and—with sufficient replication across grouping levels—decompose it into its causal constituents. Such explicit modelling is our preferred option because differences in variances can tell biological and evolutionary stories, especially when trends in variability are opposite of what is expected from Taylor's law (e.g., ceiling and floor effects can create such patterns [13]; Fig 1D).

Explicit modelling of heteroscedasticity also aligns with ongoing efforts to make biomedical science more aware of individual differences (e.g., precision or personalised medicine). Adhering to adequate sample size requirements and including both sexes in empirical research by default, will improve the generalisability of research and encourage more biologically realistic planning of future experiments. However, meaningful modelling of heteroscedasticity would still require large sample sizes for both sexes. Of relevance, sample sizes, required in the generalised least-squares framework with variance heteroscedasticity modelled explicitly, yield comparable sample size requirements to scenarios not modelling heteroscedasticity. Sandwich estimators equalise sample size requirements for low and large heteroscedasticity scenarios, but sample size requirements remain markedly higher than in homoscedastic scenarios (see

Supplementary Information, webpage: https://szymekdr.github.io/230713_power_sex_sim/ (Section 8).

Concluding remarks

Our aim of this commentary is not to highlight difficulties in including sex as a biological variable in a study, discouraging sex inclusion. Quite the opposite, we believe sex inclusion is a critical element of study design for generalisability, agreeing with Philips and colleagues. However, sex inclusion is not a free ride, unlike what Philips and colleagues suggested. We urge that funding bodies and ethics committees recognise that if they ask for sex inclusion, they will need to allow researchers to have more subjects and budget to gain robust statistical inference and, therefore, robust conclusions (cf. [14]). There is no way around this biological and statistical reality.

Supplementary Information: An HTML file containing 8 sections: Sections 1–3, different simulation scenarios; Section 4, the impact of sample size; Section 5, simulating sample size; Section 6, source of heterogeneity, Section 7, Taylor’s law; and Section 8, additional simulation scenarios (webpage: https://szymekdr.github.io/230713_power_sex_sim/ or underlying files and data can be found in: <https://doi.org/10.5281/zenodo.10205440>).

Acknowledgments

A part of the writing was conducted while visiting the Okinawa Institute of Science and Technology (OIST) through the Theoretical Sciences Visiting Program (TSVP) to SN.

Author Contributions

Conceptualization: Szymon M. Drobniak, Malgorzata Lagisz, Yefeng Yang, Shinichi Nakagawa.

Formal analysis: Szymon M. Drobniak, Shinichi Nakagawa.

Funding acquisition: Szymon M. Drobniak, Malgorzata Lagisz, Shinichi Nakagawa.

Investigation: Szymon M. Drobniak, Shinichi Nakagawa.

Methodology: Szymon M. Drobniak, Shinichi Nakagawa.

Visualization: Szymon M. Drobniak, Malgorzata Lagisz.

Writing – original draft: Szymon M. Drobniak, Shinichi Nakagawa.

Writing – review & editing: Szymon M. Drobniak, Malgorzata Lagisz, Yefeng Yang, Shinichi Nakagawa.

References

1. Zucker I, Beery AK. Males still dominate animal studies. *Nature*. 2010; 465(7299):690. <https://doi.org/10.1038/465690a> WOS:000278551800020. PMID: 20535186
2. Arnegard ME, Whitten LA, Hunter C, Clayton JA. Sex as a Biological Variable: A 5-Year Progress Report and Call to Action. *J Womens Health*. 2020; 29(6):858–864. <https://doi.org/10.1089/jwh.2019.8247> WOS:000508893800001. PMID: 31971851
3. Phillips B, Haschler TN, Karp NA. Statistical simulations show that scientists need not increase overall sample size by default when including both sexes in in vivo studies. *PLoS Biol*. 2023; 21(6):e3002129. <https://doi.org/10.1371/journal.pbio.3002129> PMID: 37289836
4. Cohen JE, Xu M. Random sampling of skewed distributions implies Taylor’s power law of fluctuation scaling. *Proc Natl Acad Sci U S A*. 2015; 112(25):7749–7754. <https://doi.org/10.1073/pnas.1503824112> WOS:000356731300068. PMID: 25852144

5. Anwer H, Morris MJ, Noble DWA, Nakagawa S, Lagisz M. Transgenerational effects of obesogenic diets in rodents: A meta-analysis. *Obes Rev.* 2022; 23(1). <https://doi.org/10.1111/obr.13342> WOS:000702130000001. PMID: 34595817
6. Wyman MJ, Rowe L. Male Bias in Distributions of Additive Genetic, Residual, and Phenotypic Variances of Shared Traits. *Am Nat.* 2014; 184(3):326–337. <https://doi.org/10.1086/677310> WOS:000340844300007. PMID: 25141142
7. Prendergast BJ, Onishi KG, Zucker I. Female mice liberated for inclusion in neuroscience and biomedical research. *Neurosci Biobehav R.* 2014; 40:1–5. <https://doi.org/10.1016/j.neubiorev.2014.01.001> WOS:000333786900001. PMID: 24456941
8. Zajitschek SRK, Zajitschek F, Bonduriansky R, Brooks RC, Cornwell W, Falster DS, et al. Sexual dimorphism in trait variability and its eco-evolutionary and statistical implications. *Elife.* 2020; 9. ARTN e63170 <https://doi.org/10.7554/eLife.63170> WOS:000595590400001. PMID: 33198888
9. Wilson LAB, Zajitschek SRK, Lagisz M, Mason J, Haselimahhadi H, Nakagawa S. Sex differences in allometry for phenotypic traits in mice indicate that females are not scaled males. *Nat Commun.* 2022;13(1). ARTN 7502 <https://doi.org/10.1038/s41467-022-35266-6> WOS:000969735000018. PMID: 36509767
10. Cleasby IR, Nakagawa S. Neglected biological patterns in the residuals A behavioural ecologist's guide to co-operating with heteroscedasticity. *Behav Ecol Sociobiol.* 2011; 65(12):2361–2372. <https://doi.org/10.1007/s00265-011-1254-7> WOS:000297120000017.
11. Zeileis A, Köll S, Graham N. Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R. *J Stat Softw.* 2020; 95(1):1–36. <https://doi.org/10.18637/jss.v095.i01> WOS:000581028400001.
12. Viechtbauer W, Lopez-Lopez JA. Location-scale models for meta-analysis. *Res Synth Methods.* 2022; 13(6):697–715. <https://doi.org/10.1002/jrsm.1562> WOS:000787789500001. PMID: 35439841
13. Yang YF, Lagisz M, Foo YZ, Noble DWA, Anwer H, Nakagawa S. Beneficial intergenerational effects of exercise on brain and cognition: a multilevel meta-analysis of mean and variance. *Biol Rev.* 2021; 96(4):1504–1527. <https://doi.org/10.1111/brv.12712> WOS:000634460200001. PMID: 33783115
14. Nakagawa S, Lagisz M, Yang Y, Drobniak S. Finding the right power balance: better study design and collaboration can reduce dependence on statistical power. 2023. <https://doi.org/10.1371/journal.pbio.3002423> PMID: 38190355