Received: 4 May 2018 | Accepted: 9 May 2018

DOI: 10.1002/ajpa.23614

LETTER TO THE EDITOR

Reply to Ruff, Warden, and Karlson

We welcome Ruff, Warden, and Karlson's letter in response to our recent (Peacock et al., 2018) analysis of the relative differences in skeletal traits among and within three strains of mice. Our goal was to highlight the challenges of interpreting skeletal phenotypes as primarily the result of habitual physical activity. We found higher prevalence of differences in skeletal phenotypes among strains of mice than we found resulting from chronic voluntary exercise within strains. Ruff and colleagues disagree with our work from both philosophical and analytic standpoints, and we respond to their concerns here.

Ruff et al. criticize our choice of mouse strains (*Myh4*^{Minimsc}, C57BL/6NHsd, and C3H/HENHsd; hereafter MM, C57, and C3H) to test for the relative effects of seven weeks of voluntary exercise on skeletal phenotypes. They make several erroneous statements regarding the nature of the strains we chose for our study. The partially inbred MM strain was derived from line HR6 of the ongoing selection experiment in the Garland lab at generation 51. In that experiment, mice in all four of the replicate High Runner (HR) lines are bred for voluntary wheel running, not muscle mass. They are never intentionally inbred. We began a separate experiment to inbreed mice from the HR lines that were homozygous for a mutation in the *Mhy4* gene (*Myh4*^{Minimsc}: Kelly et al., 2013), which results in a near lack of type IIb myosin. To create this inbred HR line 6, we began with individuals known to be homozygous for the mutation and followed standard brother-sibling mating for many generations, as described in the paper. Thus, the MM allele could not have been selected for during the process of intentional inbreeding, as Ruff and colleagues assert, because it was already fixed in the population's founding individuals. In contrast to other experiments (Cosman, Sparrow, & Rolian, 2016), the MM mice have never been artificially selected for skeletal phenotypes. Thus, any phenotypic differences are a correlated response to selection for high amounts of voluntary exercise (Garland et al., 2002).

The C57 and C3H strains were chosen for study because they had been previously shown to differ in mineralization (Beamer et al., 1996), and we sought to put MM skeletal phenotypes into a comparative context. Historically, these strains were never artificially selected for skeletal phenotypes, contrary to what Ruff et al. assert. Rather, the C57 strain was developed as a general-purpose inbred strain ("Developed in 1921 by Little from brother-sister pair (female 57 × male 52) of Miss Abby Lathrop's stock."; envigo.com), and the C3H strain was developed by selection for mammary tumors (jax.org). Only later were the high mineral apposition and bone formation rates in C3H identified as leading to skeletal phenotypic differences (Sheng et al., 1999). Any concern about inbreeding reducing the possibility of phenotypic plasticity is mitigated by the fact that the vast majority of experimental bone biomechanical research is carried out using inbred strains, and skeletal plasticity has been demonstrated numerous times.

A major source of disagreement—one leading to dramatically different conclusions—concerns our statistical approaches. Ruff et al. use frequentist approaches and analyze ratios, which involve statistical models that often produce and are sensitive to extreme values and that are prone to overfitting. Although other statistical methodologies can address overfitting, we favored a Bayesian approach using weakly regularizing priors. This approach has the benefit of automatically down-weighting observations at the extremes of the observed data by assigning them less probability. Extreme values are known to disproportionately influence frequentist inference (e.g., exert high "leverage" on regression models), which can lead to increased type I errors (Fox, 1997). Furthermore, Bayesian inference, combined with model comparison, allows for testing the relative merits of different statistical models, the consilience between model and data. For example, (1) a model which includes only a covariate, can be compared with (2) a model with inbred strain as a predictor, (3) a model which includes both strain and wheel access as predictors, and (4) a model including strain, wheel access, and their interaction term. In Bayesian model comparison, each model will receive a weighting score that is proportional to the relative support for each model, based on its expected out-of-sample predictive ability, given the data and the priors (Vehtari, Gelman, & Gabry, 2017; Yao, Vehtari, Simpson, & Gelman, 2018). Simply fitting the full factorial model with interactions, as Ruff and colleagues do, does not allow for such nuanced comparisons, and the full factorial model is very likely to overfit the data. Although for simplicity of presentation, we did not include these models in our recent paper (Peacock et al., 2018), we explored these methods before settling on a single model for analysis. Overall, we prefer the more conservative approach that Bayesian inference affords the researcher, when compared to the chimera of Fisherian arbitrary *p*-value cutoffs of 0.05 combined with Neyman-Pearson null hypothesis testing to define statistical "significance" that is currently the standard (see Smith, 2018 for additional criticism).

We attempted to but, with one exception, were unable to reproduce Ruff et al.'s statistical analysis of our data, i.e., to match their *p* values. They provided neither any code to reproduce their analyses, nor adequate detail about their statistical methodology for reproducibility. We have done our best to mimic their analyses, but in only one case were able to calculate equivalent *p* values (their Figure 3), which we find somewhat disconcerting. Our reanalysis of our data here used both frequentist linear models for closest approximation to Ruff et al.'s analysis (i.e., R function lm) and Bayesian linear models estimated using the stan sampler (Carpenter, et al. 2017), the rstanarm package (https://cran.r-project.org/package=rstanarm) for

TABLE 1 Model comparison of four nested models predicting log-transformed femoral length with combinations of log-transformed body mass, strain, and wheel access

Model: ln(femoral length) ~	LOO ELPD (se)	Model weight
ln(mass)	104.2 (4.1)	0%
ln(mass) + strain	**115.2 (4.4)**	**100%**
ln(mass) + strain + wheel	114.3 (4.3)	0%
ln(mass) + strain + wheel + strain x wheel	112.4 (4.0)	0%
ln(mass) + strain + strain x ln(mass)	113.6 (3.8)	0%

LOO ELPD is the leave-one-out expected log predictive density, a measure of predicted out-of-sample model performance (Vehtari et al., 2017), which is similar to information criteria (e.g., AIC, WAIC). Model weight is a measure of the relative compatibility between a set of models and the observed data. Model weight is distributed between a set of models according to relative compatibility. Although model weights vary continuously, in this case, 100% of the model weight falls on a model with body mass as a covariate and strain but not with wheel access. The conclusion from model comparison is that there is little credible evidence for an effect if wheel access.

posterior estimation of parameters, and the loo package (https://cran.r-project.org/package=loo) for out-of-sample model performance and model comparison. All analyses were carried out using R (ver. 3.5), and all code to reproduce our statistical analyses is included as supplementary information (https://osf.io/3wxnh/).

From an analytic standpoint, Ruff et al. disagree primarily with our use of femoral length as the covariate in our linear models in order to account for skeletal size differences between animals. They argue that body mass is a better measure of mechanical loading on the skeleton, a point with which do not disagree. However, body mass is a poor covariate for size correction in this sample. As noted in our original communication and as agreed to by Ruff and colleagues, body mass is highly sensitive to chronic exercise. In our sample of mice, body mass is credibly different both between strains and across wheel access treatments. Wheel access consistently led to a ∼4 g decrease in body mass within strains. Femoral length also was credibly different among strains, and a small but credible decrease in femoral length is associated with wheel access. The magnitudes of these difference are not equal, however. Body mass is between 16.3 and 22.6% lower in the exercise treatments. In contrast, femoral length cubed is between 3.2 and 4.8% lower in exercise treatments.

A reasonable question is then: which of these variables, when used as a covariate, more adequately controls for the covariance between size and the variable of interest? Although body mass may be a better proxy for loading via ground reaction forces, if it does not adequately account for overall differences in skeletal size, then any apparent patterns may be a statistical artifact. Based on the analysis described below, we conclude that femoral length is preferable to body mass for this sample, even though, as we show below, we draw the same conclusions using body mass.

Although we were not able to exactly reproduce the *p* values in Ruff et al.'s (2018) Figure 1, even using a factorial ANCOVA on our own data, we agree that log-transformed femoral length predicted by log-transformed body mass are credibly different between strains. The posterior intervals of differences between MM and C57 or C3H have about ∼99% posterior probability of not including zero, with no credible effect of wheel access (Table 1). We find a similar result with frequentist likelihood-ratio tests (P for including wheel access in the model is 0.7). A further criticism of the model presented in Ruff et al.'s Figure 1 is that there is no credible evidence for an interaction between body mass and strain, which the presence of strain-specific slopes in the figure implies (i.e., the model in the bottom row of our Table 1).

Use of femoral length rather than body mass is further argued for given the observation that femoral length is a trait that is able to be measured from skeletons in the fossil and archaeological records. Thus, an analysis using femoral length as a covariate is of greater potential utility to other researchers. Estimates of body mass in the fossil/archaeological record and for many museum specimens will never be more than estimates. Our model of body mass predicted by femoral length shows that for a given femur length, it is not possible to statistically distinguish between strains of mice with large differences in body mass, even when these differences are relatively large (i.e., >15%).

Ruff et al. reanalyze our data for estimates of bending rigidity (I_{ap} and I_{ml}; in Ruff et al. 2018 Figure 2). They first convert the relevant data to ratios: ln(I_{ap}/mass) and ln(I_{ap}/mass). The rationale either for this transformation or for switching from an ANCOVA model (their Figure 1) to a ratio is not made clear in their text. For decades, the undesirable statistical properties of ratios (Atchley, Gaskins, & Anderson, 1976) and in particular the inadequacy of ratios for size standardization (Albrecht, Gelvin, & Hartman 1993) have been well known. Of particular concern, ratios assume a constant relationship between the numerator and denominator and exaggerate relationships between them, and thus can bias results of subsequent analyses. As noted above, we were unable to exactly replicate Ruff et al.'s analysis of our data (i.e., *p* values in Ruff et al. 2018 Figure 2) with the information they provide in their letter. Our models that best approximate theirs have similar *p* values for some parameter estimates (see Supplemental Information; https://osf.io/3wxnh/), but this nominal agreement might result from chance alone. Given the known inadequacies of ratios, we reject the use of ratios in favor of comparing a family of linear models with both categorical predictors (strain and wheel access and their interaction) and a continuous covariate (log-transformed body mass).

We thought it would be illustrative to do what Ruff and colleagues suggest and perform the analysis of our data using body mass as a covariate in an ANCOVA-like linear model (using both frequentist and Bayesian approaches). To this end, for each of I_{ml} and I_{ap}, we compared four nested linear models of the form:

1. ln(trait) ∼ ln(covariate)
2. ln(trait) ∼ ln(covariate) + Strain
3. ln(trait) ∼ ln(covariate) + Strain + Wheel
4. ln(trait) ∼ ln(covariate) + Strain + Wheel + Strain x Wheel

where "trait" is either I_{ml} or I_{ap} (as in Table 1, but omitting the Strain × ln(Mass) model), and "covariate" is either body mass or femoral length. If Ruff et al.'s assertions are correct, then the best-supported model should be either models 3 or 4, which include strain, wheel access, and

TABLE 2 Model comparison of four nested models predicting log-transformed ML second moment of area with combinations of log-transformed body mass, strain, and wheel access

Model: ln(I_{ml}) ~	LOO ELPD (se)	Model weight
ln(mass)	−0.6 (4.1)	0%
ln(mass) + strain	**25.8 (4.4)**	**87%**
ln(mass) + strain + wheel	25.2 (4.7)	13%
ln(mass) + strain + wheel + strain x wheel	24.1 (4.1)	0%

Column descriptions follow Table 1.

possibly the strain × wheel access interaction. Results of our reanalysis are summarized in Tables 2 and 3, which are comparable to Ruff et al.'s (2018) Figure 2a and 2c (full details and models similar to Figures 2b and 2d are provided online: https://osf.io/3wxnh/). In all cases, the model with the highest weight is model 2 (83 and 90% of model weight), which includes a parameter estimate for strain but not for wheel. We find similar results using frequentist models compared with likelihood-ratio tests: in all cases, the model in not improved by including a wheel access parameter ($p = 0.32$ and 0.96). Thus, we find that, when properly analyzed using a linear model instead of ratios, and even in a frequentist framework and with body mass as a covariate, there is no evidence in support of Ruff et al.'s statistical conclusions about I_{ap} and I_{ml}. Finally, we do find a credible effect of wheel access in I_{ap}/I_{ml}, which is unsurprising, given previously observed shape differences in MM mice relative to unaffected mice (Middleton et al., 2010).

We would like to emphasize that our intention is not to argue that loading has no effect on skeletal morphology. We recognize that habitual activities can induce observable variation in bone cross-sectional properties within populations. However, it is also important to recognize that there can be baseline, population-level differences in both skeletal phenotypes and in how skeletal elements respond to exercise. This observation has been demonstrated in outbred mouse models (Wallace, Judex, & Demes, 2015), as well as in modern human populations (Meiring, Avidon, Norris, & McVeigh, 2013). Understanding that different populations can have different skeletal responses to the same habitual loading environments is necessary for the accurate interpretation of the fossil record, because it means that we must be cautious assigning any particular behavior to a given bone morphology.

TABLE 3 Model comparison of four nested models predicting log-transformed AP second moment of area with combinations of log-transformed body mass, strain, and wheel access

Model: ln(I_{ap}) ~	LOO ELPD (se)	Model weight
ln(mass)	20.3 (5.2)	10%
ln(mass) + strain	**25.3 (5.2)**	**90%**
ln(mass) + strain + wheel	24.4 (5.1)	0%
ln(mass) + strain + wheel + strain x wheel	23.5 (4.5)	0%

Column descriptions follow Table 1.

Ultimately, our original paper strives to make two points. The first is that femoral cross-sectional morphology is not always an accurate indication of resistance to bending. Despite Ruff et al.'s apparent concerns, we are confident that our experimental design did in fact allow us to test for associations between inferred and actual resistance to bending. The second, arguably more important point is that increased resistance to bending or loading does not necessarily indicate increased physical activity and that some care should be taken when interpreting cross-sectional properties as direct indicators of activity. Although we recognize that bones can and do produce observable variation in response to different loading environments, we question the assertion that observing such phenotypic differences in isolated fossil remains can give definitive information about whether the source of skeletal loading was activity-driven ground reaction or muscular forces rather than gravitational forces acting on body mass.

In conclusion, we disagree with Ruff et al.'s assertions that our study was flawed in conception and analysis. We remain confident that strain-level differences in skeletal phenotypes are likely to be more frequent and of greater magnitude than differences within strains resulting from chronic exercise.

ORCID

Sarah J Peacock http://orcid.org/0000-0002-1064-9482
Theodore Garland, Jr http://orcid.org/0000-0002-7916-3552
Kevin M Middleton http://orcid.org/0000-0003-4704-1064

Sarah J Peacock[1], Theodore Garland, Jr.[2],
Kevin M. Middleton[1]
[1]*Department of Pathology and Anatomical Sciences, University of Missouri, Columbia, Missouri*
[2]*Department of Biology, University of California, Riverside, California*

Correspondence
Sarah J. Peacock, Department of Pathology & Anatomical Sciences, University of Missouri, Medical Sciences M263, 1 Hospital Dr., Columbia, MO 65212.
Email: sjpd58@mail.missouri.edu

REFERENCES

Albrecht, G. H., Gelvin, B. R., & Hartman, S. E. (1993). Ratios as a size adjustment in morphometrics. *American Journal of Physical Anthropology*, *91*, 441–468.

Atchley, W. R., Gaskins, C., & Anderson, D. (1976). Statistical properties of ratios. I. empirical results. *Systematic Zoology*, *25*, 137–148.

Beamer, W. G., Donahue, L. R., Rosen, C. J., & Baylink, D. J. (1996). Genetic Variability in Adult Bone Density Among Inbred Strains of Mice. *Bone*, *18*, 397–403.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1–32.

Cosman, M. N., Sparrow, L. M., & Rolian, C. (2016). Changes in shape and cross-sectional geometry in the tibia of mice selectively bred for increases in relative bone length. *Journal of Anatomy*, *228*, 940–951.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: SAGE Publications, Inc.

Garland, T., Jr, Morgan, M. T., Swallow, J. G., Rhodes, J. S., Girard, I., Belter, J. G., & Carter, P. A. (2002). Evolution of a small-muscle polymorphism in lines of house mice selected for high activity levels. *Evolution: International Journal of Organic Evolution*, *56*, 1267–1275.

Kelly, S. A., Bell, T. A., Selitsky, S. R., Buus, R. J., Hua, K., Weinstock, G. M., ... Pomp, D. (2013). A novel intronic single nucleotide polymorphism in the myosin heavy polypeptide 4 gene is responsible for the mini-muscle phenotype characterized by major reduction in hind-limb muscle mass in mice. *Genetics*, *195*, 1385–1395.

Meiring, R. M., Avidon, I., Norris, S. A., & McVeigh, J. A. (2013). A two-year history of high bone loading physical activity attenuates ethnic differences in bone strength and geometry in pre-/early pubertal children from a low-middle income country. *Bone*, *57*, 522–530.

Middleton, K. M., Goldstein, B. D., Guduru, P. R., Waters, J. F., Kelly, S. A., Swartz, S. M., & Garland, T., Jr. (2010). Variation in within-bone stiffness measured by nanoindentation in mice bred for high levels of voluntary wheel running. *Journal of Anatomy*, *216*, 121–131.

Peacock, S. J., Coats, B. R., Kirkland, J. K., Tanner, C. A., Garland, T., Jr., & Middleton, K. M. (2018). Predicting the bending properties of long bones: Insights from an experimental mouse model. *American Journal of Physical Anthropology*, *165*, 457–470.

Ruff, C. B., Warden, S. J., & Carlson, K. J. (2018). Of mice and men (and women): comment on Peacock et al. 2018. American Journal of Physical Anthropology, this volume.

Sheng, M. H.-C., Baylink, D. J., Beamer, W. G., Donahue, L. R., Rosen, C. J., Lau, K.-H. W., & Wergedal, J. E. (1999). Histomorphometric studies show that bone formation and bone mineral apposition rates are greater in C3H/HeJ (high-density) than C57BL/6J (low-density) mice during growth. *Bone*, *25*, 421–429.

Smith, R. J. (2018). The continuing misuse of null hypothesis significance testing in biological anthropology. *American Journal of Physical Anthropology*, *166*, 236–245.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1433–1432.

Wallace, I. J., Judex, S., & Demes, B. (2015). Effects of load-bearing exercise on skeletal structure and mechanics differ between outbred populations of mice. *Bone*, *72*, 1–8.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions. Bayesian Anal. Available at: https://projecteuclid.org/euclid.ba/1516093227