A meta-analysis of whole-body and heart mass effect sizes from a long-term artificial selection experiment for high voluntary exercise

Nicole E. Schwartz[a,1], Theodore Garland, Jr.[a]

[a]Department of Evolution, Ecology, and Organismal Biology
University of California – Riverside, Riverside CA, USA

Present Address:
[1]Schmid College of Science and Technology, Chapman University, Orange CA, USA

Correspondence:
Nicole E. Schwartz
Schmid College of Science and Technology
Chapman University, Orange CA, USA
Phone: 909-727-7855; Email: *nschw002@ucr.edu*

Abstract

Selection experiments play an increasingly important role in comparative and evolutionary physiology. However, selection experiments can be limited by relatively low statistical power, in part because replicate line is the experimental unit for analyses of direct or correlated responses (rather than number of individuals measured). One way to increase the ability to detect correlated responses is through a meta-analysis of studies for a given trait across multiple generations. To demonstrate this, we applied meta-analytic techniques to two traits (body mass and heart ventricle mass, with body mass as a covariate) from a long-term artificial selection experiment for high voluntary wheel-running behavior. In this experiment, all 4 replicate High Runner (HR) lines reached apparent selection limits around generations 17-27, running approximately 2.5-3-fold more revolutions/day than the 4 non-selected Control (C) lines. Although both traits would also be expected to change in HR lines (relative heart size expected to increase, expected direction for body mass is less clear), the statistical significance has varied, despite repeated measurements. We compiled information from 33 unique studies and calculated a measure of effect size (Pearson's R). Our results indicate that, despite a lack of statistical significance in most generations, HR mice have evolved larger hearts and smaller bodies relative to Controls. Moreover, plateaus in effect sizes for both traits coincides with the generational range during which the selection limit for wheel-running behavior was reached. Finally, since the selection limit, absolute effect sizes for body mass and heart ventricle mass have gotten smaller (i.e., closer to 0).

Keywords: Artificial selection, Heart, Meta analysis, Voluntary exercise, Wheel running

1. Introduction

Selection experiments have taken on an increasingly important role in comparative and evolutionary physiology (Gibbs 1999; Bennett 2003; Garland 2003; Rhodes and Kawecki 2009; Swallow et al. 2009; Sadowska et al. 2015; Storz et al. 2015). Although the goals of such experiments vary widely, the general approach is to intentionally breed for high and/or low values of some trait of interest, at one level of biological organization, and then test for correlated responses at the same or other levels. Thus, selection experiments can be used to elucidate the biological underpinnings of complex traits (e.g., whole-organism metabolic rate, exercise behavior). By studying the responses to selection on complex traits, one may investigate their underling genetics and genomics (e.g., Konczal et al. 2016; Palma-Vera et al. 2022; Hillis and Garland 2023), as well as mechanistic relationships among sub-organismal traits that influence performance, behavior, and life history traits (Rhodes and Kawecki 2009; Swallow et al. 2009; Cohen et al. 2020).

Studies from a long-term artificial selection experiment for high levels of voluntary exercise (as measured by wheel-running behavior) have provided substantial insight regarding the underlying factors (e.g., motivation, ability) influencing individual differences in exercise behavior (Garland et al. 2011b). Briefly (as more information can be found within Methods below), four replicate High Runner (HR) lines have been bred for wheel running and compared with four non-selection Control (C) lines (Swallow et al. 1998; Garland 2003; Careau et al. 2013; Wallace and Garland 2016). Wheel running, especially when performed at high levels, is an energetically demanding behavior (Koteja et al. 1999; Swallow et al. 2001; Rezende et al. 2009; Copes et al. 2015) that involves all organ systems and, therefore, is likely to engender numerous correlated responses. As expected, many adult traits have been found to differ between HR and C lines (Rhodes et al. 2005; Garland et al. 2011a; Wallace and Garland 2016), including increased home-cage activity (spontaneous physical activity) when housed without access to wheels (Malisch et al. 2009; Copes et al. 2015), increased endurance (Meek et al. 2010) and maximal oxygen consumption (VO_2max) (Rezende et al. 2005; Kolb et al. 2010; Schwartz et al. 2023), increased brain size (Kolb et al. 2013; Schmill et

al. 2023), reduced body fat (Swallow et al. 2001; Girard et al. 2007), and altered circulating levels of some hormones (Girard et al. 2007; Malisch et al. 2007; Vaanholt et al. 2007; Garland et al. 2016). However, some apparent differences (e.g., body mass, relative heart mass) between HR and C lines have not reached statistical significance in all generations (see also Castro et al. 2021b on skeletal traits). Based on the positive correlation between body size and home range size or daily movement distance among species of mammals (Garland and Albuquerque 2017; Cloyed et al. 2021), one might have expected body size to increase in the HR lines (see also Djawdan 1993 on treadmill endurance). On the other hand, human marathoners are relatively small in body size. With respect to heart size, one would have expected it to increase in the HR lines, given its key role in the ability to have a high VO_2max (Poole and Erickson 2011; Hillman and Hedrick 2015; Gillooly et al. 2017). This raises an important question: are the differences truly non-significant or are they a product of limited statistical power and a high rate of Type II errors?

The ability to detect correlated responses to selection will depend on a variety of factors, including the number of generations that have elapsed and measurement error for the trait in question. A general problem for all replicated selection experiments with vertebrates is that the degrees of freedom for testing the effects of selection are related to the number of lines in the experiment, rather than the number of individuals measured (because the line is the experimental unit). Thus, studies with vertebrates rarely involve more than six or eight total lines, including non-selected control lines (but see Sadowska et al. 2008; Wone et al. 2015). Although studies of microorganisms and insects, such as *Drosophila*, can maintain a relatively large number of lines (e.g., Rauser et al. 2009; Lenski 2017), they may be limited by other factors (e.g., difficulty in phenotyping a given trait). Hence, across a variety of organismal models, the ability to detect correlated responses (and direct responses to selection) is limited by relatively low statistical power. This is particularly troublesome if the effect of selection is small or if sample sizes (within lines) are also limited (Cohen 1988; Rosenthal and Rosnow 2008; Halsey et al. 2015; Goh et al. 2016; Halsey 2019).

One way to increase the ability to detect correlated responses in selection experiments is through meta-analysis, which allows the compilation of evidence across

multiple studies of the same phenomenon (Rosenthal and Rosnow 2008; Goh et al. 2016). First conducted by Karl Pearson in (1904), and further developed by Ronald Fisher in (1948), a meta-analysis is the statistical analysis of a collection of results from individual studies for the purpose of integrating their findings (Gene Glass is often credited as the first to coin the term in 1976). This process redirects the focus from the P-values of individual studies towards overall effect sizes and collective values for statistical significance. A meta-analysis also allows one to describe the variability of effect sizes, as well as the nature of factors that may predict their relative magnitude (i.e., "moderator variables"). Thus, the "big picture" is clarified by leveraging the statistical power provided by a larger collective sample size (Rosenthal and Rosnow 2008; Goh et al. 2016). Although meta-analysis may seem an obvious choice for selection experiments that have resulted in multiple publications regarding a given potentially correlated trait, few have taken this approach (Most et al. 2011; Khan et al. 2024).

The HR mouse model is an excellent candidate for meta-analysis, as several of the drawbacks of meta-analytic procedures do not apply (many of which are addressed within Rosenthal and Rosnow 2008). For example, (a) sampling bias (e.g., the "file drawer problem"), (b) heterogeneity of methods (i.e., the "apple to oranges" issue, as described in Glass 1976), and (c) non-independence (e.g., of individuals or effect sizes). (a) The HR mouse selection experiment is ongoing (100+ generations) and has an extensive publication history (>190 publications). Although individual studies may lack sufficient power to detect small effect sizes, a sample size this large, when combined with the additional power provided by meta-analytic procedures, allows for detection of even small effect sizes. (b) Additionally, each publication uses similar methodology (a benefit of a single lab conducting repeated measurements of traits) and a near-identical statistical model for analyses. (c) Finally, a straightforward example for conducting a meta-analysis on selection experiments would be one that (i) draws from multiple selection experiments, (ii) which each used a similar model organism and the same selection criterion, and (iii) were conducted by independent researchers. Such a meta-analysis would then be able to determine the effect of selection (broadly) by calculating independent effect sizes from each of the independent selection experiments, which in

turn would be comprised of independent studies (e.g., as seen in van der Most et al. 2011). However, in the present study, effect sizes are from separate studies of the same "experiment" (i.e., the same four HR and C lines are used in each study); but a meta-analysis may draw on parameters calculated from overlapping data sets (e.g., White et al. 2007). Moreover, we have restricted our observations (and therefore, our conclusions) to the context of the HR selection experiment (i.e., the mean of the observed effect sizes provides a robust estimate of the "real" effect size from the wheel-running experiment), but not to other potential wheel-running selection experiments (which may produce alternative solutions).

In the present study, we applied meta-analytic techniques to body mass and heart (ventricle) mass (with body mass as a covariate), two traits that have been measured repeatedly and for which statistical significance has varied (see below). The objectives of this study were: (1) to provide a methodological demonstration of the utility of meta-analytic procedures for summarizing selection experiments (in particular, long-term selection experiments), and in so doing, (2) to address long-standing questions (e.g., overall statistical significance, patterns with respect to age, sex, and number of generations of selection) about the underlying trends in two of the most commonly measured characteristics within the HR selection experiment (apart from wheel-running behaviors). Finally, the present study is intended to be part of a broader synthesis of traits related to voluntary exercise, across multiple levels of biological organization, that have been studied in the HR mouse model (see also Khan et al. 2024).

2. Methods

2.1. The High Runner mouse selection experiment

The HR mouse model is a long-term artificial selection experiment for high voluntary wheel-running behavior in laboratory house mice (*Mus domesticus*), and has been ongoing since 1993 (now over 100+ generations) (Swallow et al. 1998; Garland 2003; Garland et al. 2011b). Each generation has followed a standard protocol as follows: (1) mice are weaned at 21 days of age; (2) housed four per cage from weaning until sexual maturity; (3) housed individually with access to an exercise wheel (1.12-m circumference) for 6 days beginning at ~6-8 weeks of age; (4) for the 4 replicate HR lines, the highest-running male and female from each family are chosen as breeders for the next generation (within-family selection; no sibling pairings allowed), whereas in the 4 replicate C lines, breeders are chosen without regard to wheel running; (5) males and females are co-housed for 18-days; (6) males are removed from cages on the 19th day; (7) offspring births occur, typically over the span of 1 week; the cycle starts again from point (1) with each subsequent generation.

The original base population of mice used to start the HR selection experiment included a very small fraction (~0.5%) of individuals with hindlimb muscles that were ~50% smaller than normal-muscled individuals (Garland et al. 2002; Houle-Leroy et al. 2003). The aptly named "mini-muscle phenotype" is caused by an autosomal, single-nucleotide polymorphism, located in an intron of the *Myosin heavy polypeptide 4* (*Myh4*) skeletal muscle gene, and acts as a Mendelian recessive trait (Kelly et al. 2013). The mini-muscle phenotype results in an ~50% reduction in hindlimb muscle mass, primarily from reduced Type IIb muscle fibers (Guderley et al. 2008; Bilodeau et al. 2009), along with many other pleiotropic effects (Houle-Leroy et al. 2003; Guderley et al. 2006; Copes et al. 2015; Castro et al. 2021a; Schwartz et al. 2023), including functional characteristics hypothesized to facilitate high levels of voluntary wheel-running behavior. The phenotype was only observed in one C line and two HR lines, and eventually disappeared from the C line 5, became fixed in HR line 3, and remains polymorphic in HR line 6 (lab designations).

The statistical model used for analyses has remained relatively unchanged. Generally, mixed models are implemented using SAS Proc MIXED (SAS Institute, Cary, NC, USA) with restricted maximum likelihood (REML) estimation. Linetype is a fixed effect and replicate line (4 HR and 4 C) is nested within linetype as a random effect using the containment method for d.f., such that the d.f. for linetype are always 1 and 6. When present, mini-muscle status is included as an additional fixed effect tested relative to the residual d.f. Covariates are used as appropriate.

2.2. Criteria for inclusion in meta-analysis

The HR selection experiment has an extensive publication history (available at: https://sites.google.com/ucr.edu/hrmice/publications). Unpublished studies, including the dissertations of former graduate students, were also available. Studies were included if: (1) both body and heart (ventricle) mass were available, (2) all 8 lines were sampled, and (3) a non-experimental linetype group (e.g., HR and C mice without wheels) was present. Presence of the mini-muscle phenotype, although included when available, was not a criterion for inclusion. Some generations, particularly earlier in the selection experiment, may not have any mini-muscle mice, due to the phenotype's low frequency (described above). Implementation of these criteria resulted in 33 unique sets of data. Some studies included both males and females, and were subsequently split by sex, resulting in 48 total effect sizes for each trait (25 for females, 23 for males). These were subdivided into two groups, data from before the selection limit for voluntary wheel-running behavior (~ generation 30, as per Careau et al. 2013), and data from after the limit. Thus, we had 24 estimates of effect sizes (per trait) from before generation 30 (12 for females, 12 for males) and 24 from after generation 30 (13 for females, 11 for males).

2.3. Statistical analyses

Data gathered for the present study (such as age, sex, and sample size) are reported in Supplemental Table S1. Least Squares Means (LSMs), Standard Errors

(SE), F-statistics, and P-values were recorded from each study. Some data were transformed prior to analysis (e.g., log transformed), then back-transformed for LSMs. This procedure requires the computation of 95% Confidence Limits (95% CL), rather than Standard Errors (SE); therefore, we calculated Upper and Lower 95% CL (UL and LL respectively) for all traits, regardless of whether data were transformed or not. When values were not available (e.g., unpublished studies), the original data were analyzed (as described above) in SAS Proc MIXED to generate the necessary LSMs, SEs, F-statistics, and P-values.

The F-statistic and degrees of freedom of each study were used to calculate an effect size estimate (in this case, Pearson's R) by using the following formula:

$$R = \sqrt{\frac{F}{F + d.f.}}$$

As noted above, d.f. for linetype comparisons (i.e., HR vs C lines) were always 1 and 6, while d.f. for mini-muscle comparison (i.e., mini-muscle vs "normal" individuals) were relative to the residual d.f. Therefore, effect sizes and the original P-values are directly, although not linearly, related. A positive effect size estimate indicates that HR mice have a larger value than C mice for a given trait (and vice versa); similarly, a positive effect size estimate indicates that mini-muscle mice have a larger value than normal individuals (and vice versa).

Effect sizes for linetype (i.e., HR vs C) and mini-muscle (i.e., mini- vs normal-muscle individuals) were separately analyzed using an ANCOVA in SPSS v.28, with sex as a fixed effect, and with generation and age as covariates. P-values < 0.05 were considered statistically significant. Because the statistical power to detect interactions is lower than for detecting main effects (Wahlsten 1990), interactions with a P-value < 0.10 were also considered statistically significant, as we have done in other studies (e.g., see Cadney et al. 2022; Cruden et al. 2024). Outliers were removed if the standardized residual was greater than 3 and/or the difference from the next value was greater than ~1 standard deviation (N = 2; Meek et al., 2009 males from generation 49 and retired female breeders from generation 99).

Additionally, Fisher's combined probability test (also referred to as Fisher's method) (Mosteller and Fisher 1948) was used to determine an overall P-value for body

mass and heart (ventricle) mass (with body mass as a covariate), using the following formula:

$$\chi^2 = -2 \sum_{i=1}^{k} \ln (P_i)$$

The resulting χ^2 is a cumulative test statistic, where k is the number of studies, and p_i is the P-value from each of the independent studies. The χ^2, with 2k degrees of freedom, can subsequently be used to calculate an overall P-value.

2.4. New data for body and heart ventricle mass

New data were collected to provide information on HR mice at a more recent point in the selection experiment. Retired male breeders (N = 126) were used from generation 97 of the selection experiment. These males were euthanized via CO_2 immediately following their 18-day breeding period (as described above). Mice were then weighed and dissected to determine heart ventricle mass. Retired female breeders (N = 92) from generation 99 were also sampled. These females were euthanized via CO_2 after offspring were weaned (21 days post-partum) and dissected immediately.

3. Results

3.1. Body mass before the selection limit

Before the selection limit, the average difference in body mass between HR and C mice was -0.97 g for females and -1.10 g for males (Table 1, Figures 1A and 1B). Only 3 of 24 measurements reported a statistically significant ($P < 0.05$) difference in body mass between HR and C lines, all in later generations (Supplemental Table S1, Figure 1C), yet Fisher's combined probability tests resulted in an overall P-value < 0.05 for both sexes (Table 1). ANCOVA indicated that the linetype effect size decreased across generations ($P_{Generation} < 0.0001$), with no effect of sex ($P_{Sex} = 0.7487$), no generation*sex interaction ($P_{Generation*Sex} = 0.6662$), and a trend for the absolute magnitude of linetype effect size to decrease with age ($P_{Age} = 0.0546$) (Table 2, Figure 1D).

Before the selection limit, the average difference in body mass between mini-muscle and normal individuals was -2.60 g for females and -2.91 g for males (Table 1, Figures 2A and 2B). Six of 17 measurements (7 data sets did not have any mini-muscle individuals) reported a statistically significant ($P < 0.05$) difference in body mass between mini-muscle and normal individuals (Supplemental Table S1, Figure 2C), and Fisher's combined probability tests resulted in an overall highly significant P-value for both sexes (Table 1). ANCOVA indicated that the mini-muscle effect size decreased across generations ($P_{Generation} = 0.0460$), with no effect of sex ($P_{Sex} = 0.8139$), no generation*sex interaction ($P_{Generation*Sex} = 0.7736$), and no effect of age ($P_{Age} = 0.8850$) (Table 2, Figure 2D).

3.2. Body mass after the selection limit

After the selection limit, the average difference in body mass between HR and C mice was -2.54 g for females and -3.14 g for males (Table 1, Figures 1A and 1B). Only 7 of 24 measurements reported a statistically significant ($P < 0.05$) difference in body

mass between HR and C lines (Supplemental Table S1, Figure 1C), but Fisher's tests resulted in highly significant overall P-values for both sexes (Table 1). ANCOVA indicated that the linetype effect size increased across generations ($P_{Generation}$ = 0.0140), with no effect of sex (P_{Sex} = 0.1070), but a significant generation*sex interaction ($P_{Generation*Sex}$ = 0.0634) wherein the body mass effect size for females increased more across generations than that for males, and a trend for the absolute magnitude of linetype effect size to increase with age (P_{Age} = 0.0640) (Table 2, Figure 1D).

After the selection limit, the average difference in body mass between mini-muscle and normal individuals was -1.37 g for females and -2.41 g for males (Table 1, Figures 2A and 2B). Six of 23 measurements (1 data set did not have any mini-muscle individuals) reported a significant (P < 0.05) difference in body mass between mini-muscle and normal individuals (Supplemental Table S1, Figure 2C); Fisher's combined probability tests resulted in an overall highly significant P-value for both sexes (Table 1). ANCOVA indicated that the mini-muscle effect size tended to increase across generations ($P_{Generation}$ = 0.0772), with a significant effect of sex (P_{Sex} = 0.0242), a significant generation*sex interaction ($P_{Generation*Sex}$ = 0.0265) wherein the body mass effect size for females increased less across generation than that for males, and no effect of age (P_{Age} = 0.2705) (Table 2, Figure 2D).

3.3. Heart mass before the selection limit

Before the selection limit, the average difference in body-mass adjusted heart ventricle mass between HR and C mice was 2.26 mg for females and 2.24 mg for males (Table 1, Figures 3A and 3B). None of 24 measurements reported a statistically significant (P < 0.05) difference in heart ventricle mass between HR and C lines (Supplemental Table S1, Figure 3C), yet Fisher's combined probability tests resulted in small overall P-values for both sexes (Table 2). ANCOVA indicated that the linetype effect size tended to increase across generations ($P_{Generation}$ = 0.0822), with no effect of sex (P_{Sex} = 0.9517), no generation*sex interaction ($P_{Generation*Sex}$ = 0.3539), and no effect of age (P_{Age} = 0.2905) (Table 2, Figure 3D).

Before the selection limit, the average difference in heart ventricle mass between mini-muscle and normal individuals was 11.50 mg for females and 15.69 mg for males (Table 1, Figures 4A and 4B). Eleven of 17 measurements (7 data sets did not have any mini-muscle individuals) reported a statistically significant ($P < 0.05$) difference in heart ventricle mass between mini-muscle and normal individuals (Supplemental Table S1, Figure 4C); Fisher's combined probability tests resulted in an overall P-value < 0.05 for both sexes (Table 1). ANCOVA indicated that the mini-muscle effect size did not significantly change across generations ($P_{Generation} = 0.6427$), with a trend for females to have smaller effect size estimates ($P_{Sex} = 0.0612$) and a significant generation*sex interaction ($P_{Generation*Sex} = 0.0885$) wherein the heart ventricle mass effect size for females increased across generation ($P_{Age} = 0.5576$) (Table 2, Figure 4D).

3.4. Heart mass after the selection limit

After the selection limit, the average difference in body-mass adjusted heart ventricle mass between HR and C mice was 7.20 mg for females and 9.51 mg for males (Table 1, Figures 3A and 3B). Five of 22 measurements reported a statistically significant ($P < 0.05$) difference in heart ventricle mass between HR and C lines (Supplemental Table S1, Figure 3C); however, Fisher's combined probability tests resulted in highly significant P-values for both sexes (Table 1). ANCOVA indicated that the linetype effect size tended to decrease across generations ($P_{Generation} = 0.0519$), with no effect of sex ($P_{Sex} = 0.1508$), no generation*sex interaction ($P_{Generation*Sex} = 0.8364$), and no effect of age ($P_{Age} = 0.1154$) (Table 2, Figure 3D).

After the selection limit, the average difference in heart ventricle mass between mini-muscle and normal individuals was 9.06 mg for females and 13.41 mg for males (Table 1, Figures 4A and 4B). Only 7 of 21 measurements (3 data sets did not have any mini-muscle individuals) reported a statistically significant ($P < 0.05$) difference in heart ventricle mass between mini-muscle and normal individuals (Supplemental Table S1, Figure 4C), but Fisher's combined probability tests resulted in highly significant

overall P-values for both sexes (Table 2). ANCOVA indicated that mini-muscle effect size did not significantly change across generations ($P_{Generation}$ = 0.9824), a trend for females to have smaller effect size estimates (P_{Sex} = 0.0782), no generation*sex interaction ($P_{Generation*Sex}$ = 0.6334), and a trend for older mice to have more positive effect size estimates (i.e., magnitude of effect size tended to increase with age) (P_{Age} = 0.0870) (Table 2, Figure 4D).

4. Discussion

4.1. Overall meta-analytic results for body size and relative heart mass in HR mice

One objective of the present study was to demonstrate the utility of meta-analytic procedures in reviewing and summarizing results from experimental evolution studies. We also proposed that a quantitative approach (e.g., here, an ANCOVA of effect sizes), as opposed to a qualitative one (e.g., do studies fall above/below nominal significance of P = 0.05), would allow us to better summarize results from the HR mouse model, and reveal underlying trends that may not be apparent when examining P-values alone. For example, if had we focused on the statistical significance of studies, only 10 of 48 measurements (3 before the selection limit, 7 after) were statistically significant for body mass and only 5 of 48 measurements (none before the selection limit, 5 after) were significant for heart mass (corrected for body mass) (Supplemental Table S1). Therefore, we may have reasonably (but incorrectly) concluded that selection for voluntary wheel-running behavior had *not* resulted in "significant" changes to either trait. However, by using a meta-analytic approach, we were able to demonstrate that: (1) absolute effect sizes for body mass and relative heart mass increased (HR mice < C for body mass; HR mice > C for heart mass) before HR mice reached a selection limit for voluntary wheel-running behavior ($P_{Generation} < 0.0001$ for body mass; $P_{Generation} = 0.0822$ for heart mass; Table 2, Figures 1 and 3); (2) this trend did not differ by sex ($P_{Generation*Sex} = 0.6662$ for body mass; $P_{Generation*Sex} = 0.3539$ for heart mass; Table 2); (3) a plateau in effect sizes for both traits coincides with the generational range during which the apparent selection limits were reached (Figures 1 and 3); (4) absolute effect sizes for both body and heart mass have gotten smaller (i.e., closer to 0) since the selection limit ($P_{Generation} = 0.0140$ for body mass; $P_{Generation} = 0.0519$ for heart mass; Table 2, Figures 1 and 3); and (5) this trend differed by sex for body mass ($P_{Generation*Sex} = 0.0634$) but not for heart mass ($P_{Generation*Sex} = 0.8364$) (Table 2).

4.2. Functional and evolutionary implications of Linetype comparisons

Prior studies have indicated that neither exhausted additive genetic variance nor counterposing natural selection related to female reproductive success (Girard et al. 2002; Keeney 2011; Careau et al. 2013) are responsible for the selection limits observed in all four of the selectively bred HR lines of mice. As a complement to the traditional quantitative-genetic explanations for selection limits, a number of studies have sought potential functional limitations either on the ability or willingness (motivation) to engage in voluntary exercise (e.g., see Rezende et al. 2006; Belke and Garland 2007; Kolb et al. 2010; Meek et al. 2010; Claghorn et al. 2016; Garland et al. 2017; Castro et al. 2024; Khan et al. 2024). Our results indicate that, despite a lack of statistical significance in many generations, HR mice have evolved to be smaller (as first reported in Swallow et al. 1999) and to have larger hearts (relative to body mass) as compared with C mice (Table 2, Figures 1 and 3).

Moreover, a plateau in effect sizes for both traits coincides with the generational range which the selection for wheel-running behavior was reached. These correlated responses in body and relative heart (ventricle) mass indicate the presence of genetic correlations with wheel-running behavior, which may have imposed constraints on the evolution of wheel running (e.g., see Weber 1990; Garland and Carter 1994; Marchini et al. 2014; Agrawal 2020). Of course, correlation does not prove causation. Rather than being causally related to wheel running, the changes in both body mass and heart mass could instead be functions of some other, currently unknown trait that is causally related to wheel running. Causality could be probed experimentally by selecting for body size, relative heart size or some other trait for which evidence suggests a causal relationship with wheel-running behavior (Garland 2003). Also, as discussed previously (Swallow et al. 1999, 2009), one could test whether one (or both) traits constrained the evolution of wheel running by selecting on both traits (e.g., body mass and wheel running) simultaneously (e.g., see Wone et al. 2015).

Importantly, absolute effect sizes for both body and heart mass have become smaller (i.e., closer to 0), particularly in the last 10-20 generations. Interestingly, this has not been accompanied by any obvious change in voluntary wheel running, as HR mice have consistently run approximately 2.5-3-fold more revolutions per day than C mice since the selection limit (i.e., over ~70 generations). If body size and/or heart size

are indeed causally related to wheel running, then this pattern of decreasing effect sizes after a selection limit, which might be viewed as a functional deterioration, implies that another trait or traits related to exercise engagement (e.g., see Section 4.4) has evolved in a compensatory fashion after the selection limit. For example, perhaps maximal heart rate increased in the HR lines as the difference in heart size from the C lines diminished (for whatever reason, e.g., inbreeding depression), thus "picking up the slack." Other studies on these lines of mice indicate that some traits that seemed to represent clear adaptations for wheel running in earlier generations, such as larger femoral heads, no longer differ between HR and C lines in later generations (Castro and Garland 2018; Castro et al. 2021b). Moreover, at the genomic level, signatures of selection differ substantially between generations 22 and 61, suggesting a phenomenon termed "genetic churn" by Hillis et al. (2024), which reflects the fact that studies of adaption across time must deal with a potentially moving target. This idea of continuing coadaptation in some traits as other falter is reminiscent of the Red Queen Hypothesis (Van Valen 1973; Langerhans 2008), wherein species involved in coevolutionary interactions must continuously “run” (i.e., evolve) to stay in the same “place” (i.e., survive) (see also Rice and Holland 1997; Sinervo and Svensson 2002).

4.3. Implications of mini-muscle effect sizes

The mini-muscle phenotype is caused by an intronic single nucleotide polymorphism, which results in a Mendelian recessive trait (Kelly et al. 2013). This phenotype was present at a frequency of ~7% in the original base population used to start the selection experiment , but has only ever been observed in one C line (C line 5, lab designation) and in two HR lines (HR line 3 and HR line 6) (Garland et al. 2002). The phenotype is no longer present in the C line, became fixed in HR line 3, and remains polymorphic in HR line 6 after > 100 generations of selection (Hiramatsu et al. 2017; Cadney et al. 2021; Castro et al. 2021a). As the name suggests, mini-muscle mice are characterized by having ~50% of the hindlimb muscle mass of normal-muscle individuals, due primarily to a systemic reduction in Type IIb muscle fibers (Guderley et al. 2008; Bilodeau et al. 2009; Talmadge et al. 2014 and references therein).

The mini-muscle phenotype is associated with several other differences in morphology, physiology, and even behavior (Garland et al. 2002; Houle-Leroy et al. 2003; Guderley et al. 2006; Kelly et al. 2013; Copes et al. 2015; Castro et al. 2021b; Schwartz et al. 2023; Castro et al. 2024; Khan et al. 2024), but the adaptive significance of these pleiotropic effects has been unclear. Regarding the present study, the mini-muscle effect size for heart mass has remained relatively consistent across generations ($P_{Generation}$ = 0.6427 before selection limit; $P_{Generation}$ = 0.9824 after selection limit), with mini-muscle mice having larger hearts than normal mice, which may be one of the effects that has caused them to be favored by the selection protocol (Garland et al. 2002). However, the mini-muscle effect size for body mass exhibits an unusual pattern, and requires a holistic examination of the available information for an accurate interpretation of the mini-muscle phenotype across generations. To review: (a) the trend for mini-muscle body mass across generations is similar to HR mice overall, i.e., decreases before the selection limit, reaches a plateau after the selection limit, and begins to increase in more recent generations (Figure 1A and Figure 2A), (b) but the trend in mini-muscle effect size differs from the linetype effect size in that (i) mini-muscle effect size is greater than 0 after the selection limit and (ii) mini-muscle effect size decreases (again) after the selection limit (Figure 1D and Figure 2D). Although these patterns are statistically significant ($P_{Generation}$ = 0.0460 before selection limit; $P_{Generation*Sex}$ = 0.0265 after selection limit), the most parsimonious explanation for this phenomenon is mathematical in nature (rather than biological); that is, the loss of mini-muscle phenotype in C line 5 and fixation of the mini-muscle phenotype in HR line 3 resulted in mini-muscle becoming confounded with linetype (rather than nested within line or crossed by linetype). This phenomenon would not be apparent (or relevant) if mini-muscle effect size were consistent across generations (e.g., as with heart mass). We note here that although the mini-muscle phenotype is part of the overall HR phenotype (in that mini-muscle mice are exclusively HR mice after the initial 25-30 generations), mini-muscle individuals often express somewhat different phenotypes for a given trait as compared with non-mini HR mice. Furthermore, the “mini-muscle phenotype” (at the level of the muscle and also associated traits) can differ somewhat between HR line 3 and HR line 6 (e.g., see Guderley et al. 2008; Bilodeau et al. 2009;

Schwartz et al. 2023). The mini-muscle phenotype in general has been suggested as an example "multiple solutions" in response to a given type of selection (Garland et al. 2011a; Castro et al. 2024). However, analyses at the level of lines (wherein HR line 6 is subdivided by mini-muscle phenotype) have been infrequent (and not always relevant to the study at hand). Future meta-analyses (e.g., of all studies presenting information on the mini-muscle phenotype) should carefully consider the aforementioned points when reviewing and summarizing key traits associated with the mini-muscle phenotype to elucidate the potential adaptive significance of trends in mini-muscle effect sizes.

4.4. Notes on effect sizes, P-values, and statistical power

Measures of effect sizes, such as Pearson's R used here, provide information on both the magnitude and direction of an effect, whereas P-values can only inform as to the probability that an effect exists. Consistency in the magnitude and direction of an effect size across multiple studies strengthens one’s conclusions. Although this is also somewhat true of P-values, they are known to both (1) consistently demonstrate statistically significant levels when sample sizes are sufficiently large, even if effects are relatively non-existent (e.g., see Bartolucci et al. 2011), and (2) fail to reach significance levels when effects are relatively small (Rosenthal and Rosnow 2008; Sullivan and Feinn 2012; Halsey et al. 2015; Wasserstein and Lazar 2016; Halsey 2019). These phenomena are due to Type I (rejecting the null hypothesis when it is true) and Type II errors (accepting the null hypothesis when it is false), respectively.

Regarding (1), statistical significance represents the probability that an observed difference between samples from different groups is due to sampling error. Typically, if $P > 0.05$, differences are assumed to be adequately explained by sampling variability. However, larger sample sizes are inherently more likely to resemble the overall population, and very large sample sizes substantially decrease the magnitude of sampling variability. Therefore, very large samples would yield statistically significant results, even if there were little to no difference between groups (e.g., as seen in Bartolucci et al. 2011). However, this is not an issue for linetype comparisons in the HR mouse model, because d.f. are small and constant (e.g., see the next paragraph).

Regarding (2), statistical power (i.e., one minus the Type II error rate) represents the sensitivity of a test to detecting an effect when one is present. Power is influenced by both sample size and effect size. When effect sizes are large, one can use a relatively small sample size and still sufficiently detect existing differences between groups (e.g., see Figure 2 in Serdar et al. 2021). Conversely, when effect sizes are small, one must have relatively larger sample sizes to sufficiently detect existing differences between groups. Thus, P-values are somewhat confounded by their relative dependence on sample sizes, whereas effect sizes are relatively insulated from such effects.

In the HR mouse model, two factors that affect the statistical power for linetype comparisons (i.e., the average values for the 4 HR vs. 4 C lines) are: (a) the number of lines (restricted to 8), which limits power, and (b) the number of mice measured per line, which influences the ability to determine the true line means. Regarding (a), mini-muscle status (described in Section 2.1) is somewhat confounded with linetype, especially in later generations, as described in Castro et al. (2021b; see also Hillis et al. 2020; Hillis and Garland 2023). Castro et al. (2021b) demonstrated that the Type I error rate for mini-muscle effects is close to the expected 5% for $\alpha = 0.05$; however, the Type I error rate for linetype effects was only 1.4%, indicating a reduced capacity to detect linetype differences within the HR mouse model using this statistical model employed here and in other studies of these lines. Regarding (b), increasing amounts of among-line variance decrease statistical power to detect linetype differences by both decreasing the power to detect differences between the average of the 4 HR and the 4 C lines, but also by increasing the overlap between the range of values within HR and C linetypes (akin to extending a folding fan within each treatment group). During earlier generations, this was not much of an issue, as neither random genetic drift nor multiple adaptive responses ("multiple solutions," e.g., see Garland et al. 2011a; Hannon et al. 2011; Hillis and Garland 2023) would yet have had much effect.

Another consequence of low statistical power is the overestimation of effect sizes (e.g., see Button et al. 2013). This may explain why effect sizes in the present meta-analysis are relatively large (e.g., R values of 0.1, 0.3, and 0.5 are traditionally categorized as small, medium, and large effect sizes, respectively Cohen 1988; Lakens 2013), despite what amounts to modest differences in body size and heart mass (e.g.,

differences in heart mass after the selection limit between HR and C females is 7.20 mg, which amounts to ~5.50% of average female heart size, Table 1). Additionally, effect sizes were calculated from the F-statistics and degrees of freedom (see Section 2.3); however, given that HR and C lines were always compared with 1 and 6 d.f., this resulted in a direct, but non-linear, relationship between R- and P-values. This non-linear relationship resulted in two phenomena: (1) effect sizes less than about 0.7 did not result in statistically significant ($P < 0.05$) differences and (2) larger effect sizes did not result in proportionally smaller P-values. All that being said, the relative consistency of the observed trends in effect size, as well as Fisher's combined statistical significance being well below nominal significance (e.g., $P = 0.0016$ for the effect sizes of female heart mass after the selection limit, Table 1), support the general trends outlined in Section 4.1.

4.5. Conclusions and future directions

Overall, our results demonstrate the utility of applying meta-analytic techniques to long-term selection experiments, in particular, regarding how meta-analyses can be used to reveal previously undiscovered trends in existing data. Going forward, this meta-analysis is intended to be part of a broader synthesis of traits related to voluntary exercise, across multiple levels of biological organization, that have been studied in the HR mouse model (see also Khan et al. 2024). Although few traits have been repeatedly measured as extensively as body and heart (ventricle) mass, some have been measured multiple times across the 100+ generations of selection (e.g., liver mass, VO_2max). These traits have also varied in their statistical significance across generations, making them good candidates for meta-analytic research. Like body and heart mass, underlying trends in the data may not be apparent on a case-by-case basis, or even by evaluation on the basis of statistical significance. Furthermore, incorporating such traits into a broader meta-analysis would allow one to determine whether the evolution of high activity behavior has coincided with broad changes in the effect sizes of multiple (potentially correlated) traits and/or if other effect sizes have changed as wheel-running behavior evolved. Additionally, a broader meta-analysis could consider

other potential moderating factors, such as seasonality (see Careau et al. 2013 on seasonality in wheel running). Finally, more information from later generations would also allow us to confirm the observed trends in the magnitude of effect sizes for body mass and heart mass after the selection limit, for which the effect sizes seem to be decreasing.

Acknowledgements

We want to thank the past and present members of the Garland lab at the University of California, Riverside. Their collective studies of the High Runner mouse model is the foundation for our work. We also thank Rosemary Presburger and Lucas Piniero for their assistance in collecting new data on body and heart ventricle mass. Finally, we thank Dr. Robert Rosenthal for many helpful discussions (recently deceased at the age of 90). As a co-founder of modern meta-analysis (along with Gene Glass), his teaching, mentorship, and enthusiasm were a strong inspiration for our meta-analytic research.

Competing Interests

No competing interests declared.

Funding

Supported by NSF grant IOS-2038528.

Data Availability

All relevant data can be found within the article and its supplementary information.

Tables and Figures

Table 1. Summary statistics for Linetype and Mini-muscle whole-body and heart (ventricle) mass (with body mass as a covariate).

		Body mass at dissection																	
		Before Selection Limit									After Selection Limit								
		Mean	Std. Dev.	Difference	R	95% LL	95%UL	X^2	d.f.	P	Mean	Std. Dev.	Difference	R	95% LL	95%UL	X^2	d.f.	P
Linetype	High Runner Females	29.78	2.32	-0.97	-0.2655	-0.3797	-0.1513	26.30	24	0.0495	26.42	3.65	-2.54	-0.5402	-0.6483	-0.4322	61.99	26	< 0.0001
	Control Females	30.74	1.87								28.97	4.06							
	High Runner Males	35.67	2.32	-1.10	-0.2403	-0.3545	-0.1261	32.78	24	0.0219	30.38	2.61	-3.14	-0.5934	-0.7111	-0.4758	45.76	22	0.0006
	Control Males	36.77	1.31								33.52	3.80							
Mini-Muscle	Mini-Muscle Females	28.58	2.66	-2.60	-0.2173	-0.3086	-0.1261	57.86	20	< 0.0001	27.00	3.58	-1.37	-0.1745	-0.2949	-0.0540	67.59	26	< 0.0001
	Normal Females	31.18	1.52								28.37	4.23							
	Mini-Muscle Males	33.85	1.86	-2.91	-0.2011	-0.3110	-0.0913	37.78	14	0.0002	30.56	4.12	-2.41	-0.2141	-0.3519	-0.0763	65.79	20	< 0.0001
	Normal Males	36.76	1.41								32.97	2.62							
		Heart ventricle mass with body mass as a covariate																	
		Before Selection Limit									After Selection Limit								
		Mean	Std. Dev.	Difference	R	95% LL	95%UL	X^2	d.f.	P	Mean	Std. Dev.	Difference	R	95% LL	95%UL	X^2	d.f.	P
Linetype	High Runner Females	125.40	7.72	2.26	0.2069	0.0174	0.3964	26.24	24	0.0498	134.50	15.64	7.20	0.4876	0.3894	0.5858	44.97	24	0.0016
	Control Females	123.14	6.18								127.29	16.09							
	High Runner Males	147.27	7.02	2.43	0.1990	0.0095	0.3885	20.81	24	0.0587	150.96	20.00	9.51	0.5938	0.4859	0.7018	41.98	20	0.0008
	Control Males	144.84	7.64								141.46	18.04							
Mini-Muscle	Mini-Muscle Females	131.19	8.56	11.50	0.2668	0.1916	0.3419	69.74	20	< 0.0001	135.45	18.18	9.06	0.2268	0.1495	0.3041	57.74	24	< 0.0001
	Normal Females	119.69	6.12								126.38	13.73							
	Mini-Muscle Males	158.31	8.32	15.69	0.3132	0.2227	0.4037	62.53	14	< 0.0001	150.71	20.22	13.41	0.3345	0.2447	0.4244	74.41	18	< 0.0001
	Normal Males	142.62	5.02								137.31	17.51							

Measurements from Supplemental Table S1 were partitioned into four groups by sex (females vs males) and by whether studies occurred before or after the selection limit for voluntary wheel-running behavior (~ generation 30, as per Careau et al. 2013). Descriptive statistics (e.g., mean, standard deviation) were calculated using SPSS v28. Values for body mass and relative heart mass are presented separately for High Runner and Control lines, mini-muscle and normal-muscle individuals, and for females and males. Effect sizes (Pearson's R) are presented as Estimated Marginal Means (EMM) with associated 95% Confidence Limits (LL = Lower Limit and UL = Upper Limit). The average difference and effect sizes are listed such that a positive value indicates HR lines (or mini-muscle individuals) have higher values for a given trait (and vice versa). Fisher's combined probability test (also referred to as Fisher's method) was used to determine an overall P-value for body size and heart mass of each group. The formula used was $\chi^2 = -2\sum_{i=1}^{k}\ln(p_i)$, where χ^2 is the cumulative test statistic, k is the number of studies, and p_i are the P-values from each of the independent studies. This formula yields a chi-squared value, with 2k degrees of freedom, which can subsequently be used to calculate an overall P-value. Statistical significance was evaluated a $P < 0.05$.

Table 2. Results for Linetype and Mini-muscle whole-body and heart mass (with body mass as a covariate) effect sizes from ANCOVA with sex, generation, and age.

		Body mass at dissection											
		Before Selection Limit						After Selection Limit					
		η_p^2	F	P	Estimate	95% LL	95% UL	η_p^2	F	P	Estimate	95% LL	95% UL
Linetype	Age	0.1725	4.17	0.0546	0.0063	-0.0001	0.0127	0.1691	3.87	0.0640	-0.0023	-0.0047	0.0001
	Generation	0.7509	60.28	< 0.0001	-0.0475	-0.0603	-0.0347	0.2782	7.32	0.0140	0.0015	-0.0053	0.0083
	Sex	0.0052	0.11	0.7487	-0.0252	-0.1872	0.1367	0.1309	2.86	0.1070	-0.4434	-0.9919	0.1051
	Generation*Sex	P = 0.6662, removed from model						0.1698	3.89	0.0634	0.0077	-0.0005	0.0159
Mini-Muscle	Age	0.0017	0.02	0.8850	0.0003	-0.0044	0.0050	0.0670	1.29	0.2705	0.0014	-0.0012	0.0040
	Generation	0.2725	4.87	0.0460	-0.0120	-0.0238	-0.0003	0.1634	3.51	0.0772	0.0011	-0.0062	0.0084
	Sex	0.0044	0.06	0.8139	-0.0162	-0.1620	0.1295	0.2516	6.05	0.0242	0.6927	0.1012	1.2842
	Generation*Sex	P = 0.7736, removed from model						0.2450	5.84	0.0265	-0.0100	-0.0188	-0.0013
		Heart ventricle mass with body mass as a covariate											
		Before Selection Limit						After Selection Limit					
		η_p^2	F	P	Estimate	95% LL	95% UL	η_p^2	F	P	Estimate	95% LL	95% UL
Linetype	Age	0.0557	1.18	0.2905	-0.0055	-0.0161	0.0051	0.1320	2.74	0.1154	-0.0018	-0.0040	0.0005
	Generation	0.1434	3.35	0.0822	0.0186	-0.0026	0.0398	0.1940	4.33	0.0519	-0.0039	-0.0078	0.0000
	Sex	0.0002	0.00	0.9517	0.0079	-0.2608	0.2766	0.1112	2.25	0.1508	-0.1062	-0.2549	0.0425
	Generation*Sex	P = 0.3539, removed from model						P = 0.8364, removed from model					
Mini-Muscle	Age	0.0294	0.36	0.5576	0.0010	-0.0026	0.0046	0.1625	3.30	0.0870	0.0015	-0.0002	0.0033
	Generation	0.0185	0.23	0.6427	-0.0058	-0.0217	0.0101	0.0000	0.00	0.9824	0.0000	-0.0031	0.0031
	Sex	0.2622	4.26	0.0612	-0.2392	-0.4915	0.0132	0.1713	3.51	0.0782	-0.1078	-0.2291	0.0135
	Generation*Sex	0.2226	3.44	0.0885	0.0158	-0.0028	0.0345	P = 0.6334, removed from model					

Measurements from Supplemental Table S1 were partitioned into four groups by sex (females vs males) and by whether studies occurred before or after the selection limit for voluntary wheel-running behavior (~ generation 30, as per Careau et al. 2013). Effect sizes for body mass and relative heart mass were analyzed using an ANCOVA in SPSS v28, with sex as a fixed effect, and generation and age as covariates. Partial eta-squared (η_p^2, an estimate of effect size for ANOVAs), F-statistics, and P-values from ANCOVAs are presented for generation, age, and sex. ANCOVAs were initially run with a Generation*Sex interaction, which was subsequently removed if not statistically significant (evaluated at $P < 0.10$; see section 2.3. Statistical analyses). Parameter estimates and associated 95% CL from ANCOVAs are presented for generation, age, sex, and the generation*sex interaction (if present).

Figure 1. Values for body mass based on linetype (High Runner vs Control). (A) Least Squares Means (LSMs) (from analyses in SAS Proc. MIXED) for body mass are from the individual studies used in the present meta-analysis. High Runner lines are depicted in red and Control lines are depicted in blue. Females are denoted by circles, males by squares. Gray is used for comparisons of HR vs C (e.g., differences, P-values, effect sizes). The dashed line indicates the selection limit for high voluntary wheel-running behavior, which occurred at approximately generation 30 (Careau et al. 2013). (B) Differences in the LSMs for body mass between HR and C lines. A positive value indicates that HR lines have higher body mass than C lines. (C) Original P-values from the individual studies used in the present meta-analysis. Dashed line is set to P = 0.05. Note that there are few studies that reach this level of statistical significance. (D) Effect sizes from the individual studies, calculated from the resultant F-statistic and degrees of freedom (generally 1 and 6 for linetype comparisons) for analyses in SAS Proc. MIXED. Effect sizes are calculated using $R = \sqrt{\frac{F}{F+d.f.}}$. A positive effect size indicates that HR lines have a larger value for body mass (i.e., that selection for high voluntary wheel-running behavior has had a positive effect on body mass). Effect sizes in the present study may be relatively large, despite somewhat small differences in body mass (e.g., differences in body mass for females after the selection limit is 2.54 g, which amounts to ~9.17% of average female body mass, Table 2), due to relatively low statistical power (as outlined in Section 4.2). That said, the relative consistency of the observed trends in effect sizes strengthens the conclusions drawn in Section 4.1.

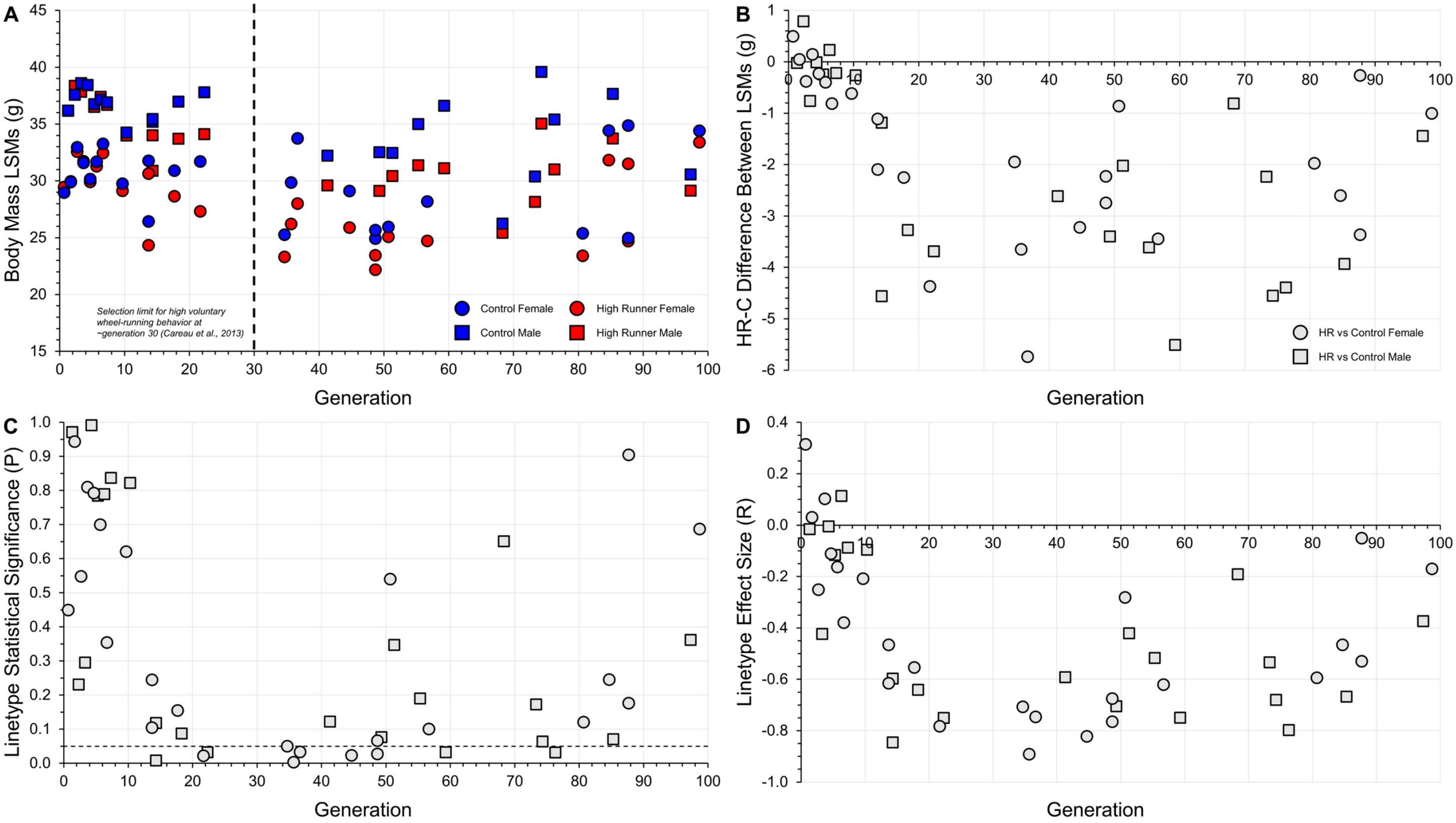
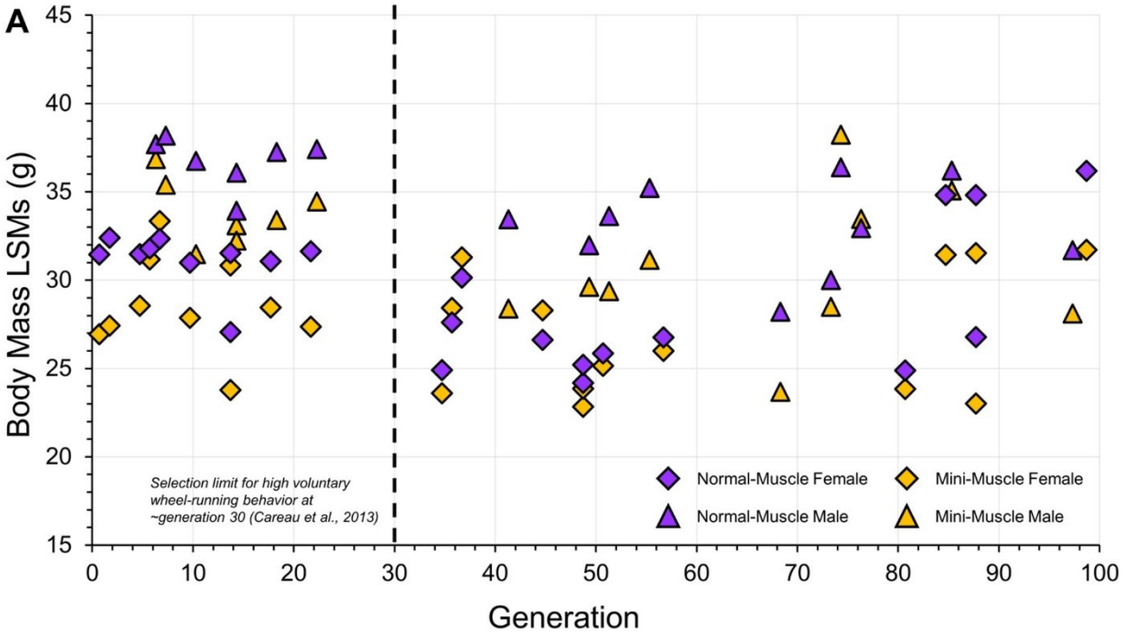
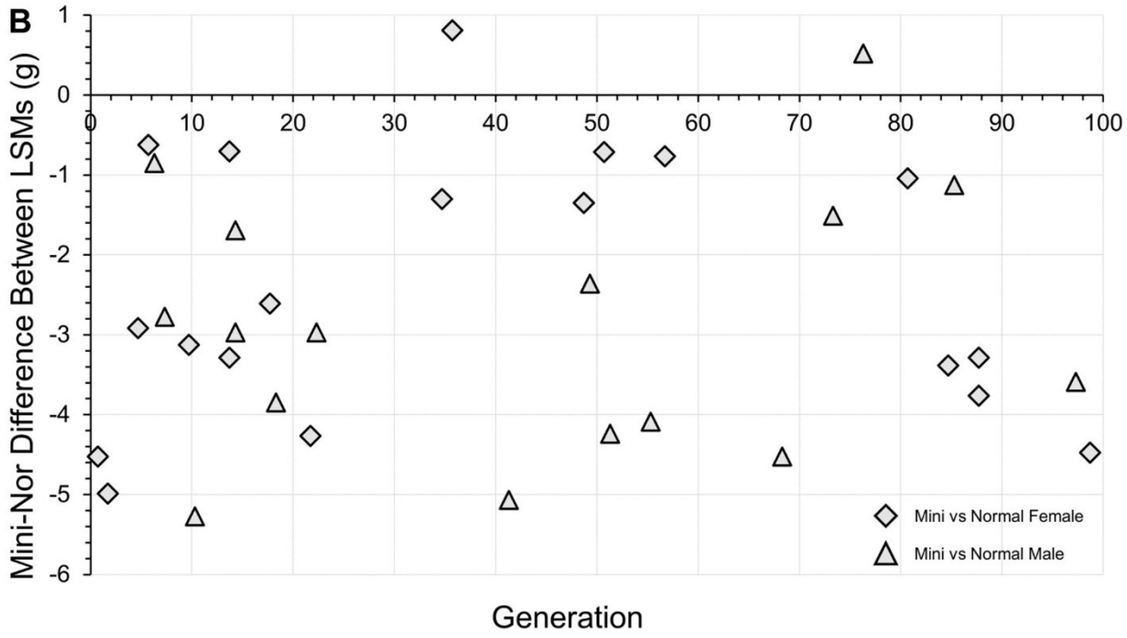
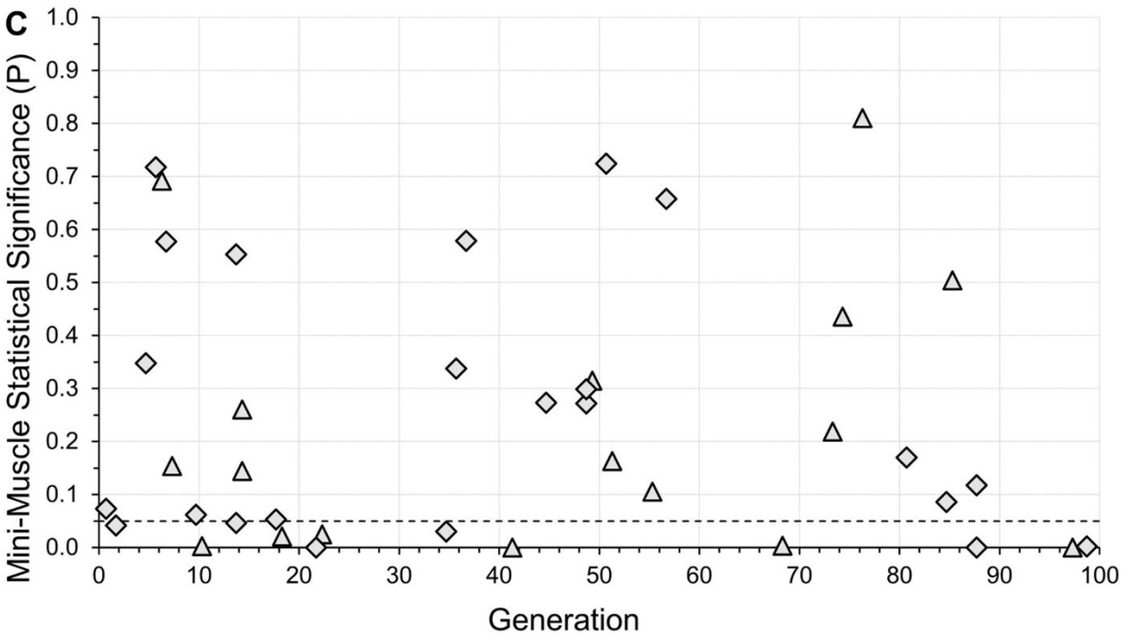

A
Body Mass LSMs (g)
Generation
Selection limit for high voluntary wheel-running behavior at ~generation 30 (Careau et al., 2013)
Control Female
High Runner Female
Control Male
High Runner Male
B
HR-C Difference Between LSMs (g)
Generation
HR vs Control Female
HR vs Control Male
C
Linetype Statistical Significance (P)
Generation
D
Linetype Effect Size (R)
Generation

Figure 2. Values for body mass based on mini-muscle status (mini-muscle vs normal-muscle individuals). Mini-muscle mice are depicted in orange and normal-muscle mice are depicted in purple. Females are denoted by diamonds, males by triangles. Gray is also used for comparisons of mini-muscle vs normal (e.g., differences, P-values, effect sizes). The dashed line indicates the selection limit for high voluntary wheel-running behavior, which occurred at approximately generation 30 (Careau et al. 2013). For more information, see descriptions above in Figure 1.


A
Body Mass LSMs (g)
Generation
Selection limit for high voluntary wheel-running behavior at ~generation 30 (Careau et al., 2013)
Normal-Muscle Female
Mini-Muscle Female
Normal-Muscle Male
Mini-Muscle Male


B
Mini-Nor Difference Between LSMs (g)
Generation
Mini vs Normal Female
Mini vs Normal Male


C
Mini-Muscle Statistical Significance (P)
Generation

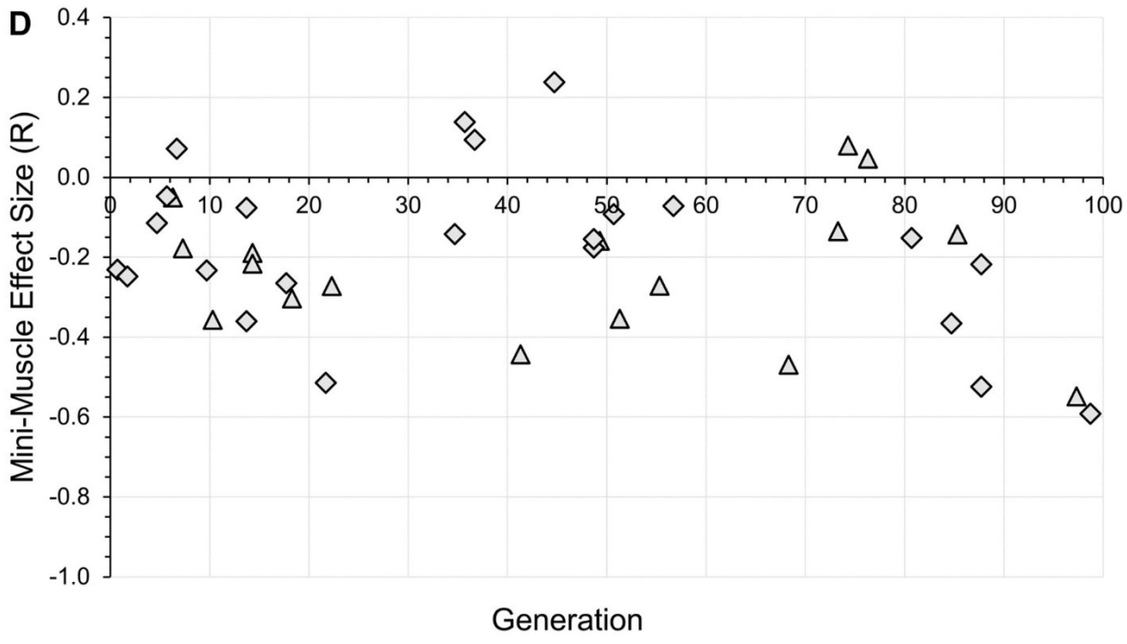

D
Mini-Muscle Effect Size (R)
Generation

Figure 3. Values for heart (ventricle) mass (with body mass as a covariate) based on linetype (High Runner vs Control). For more information, see descriptions in Figure 1.

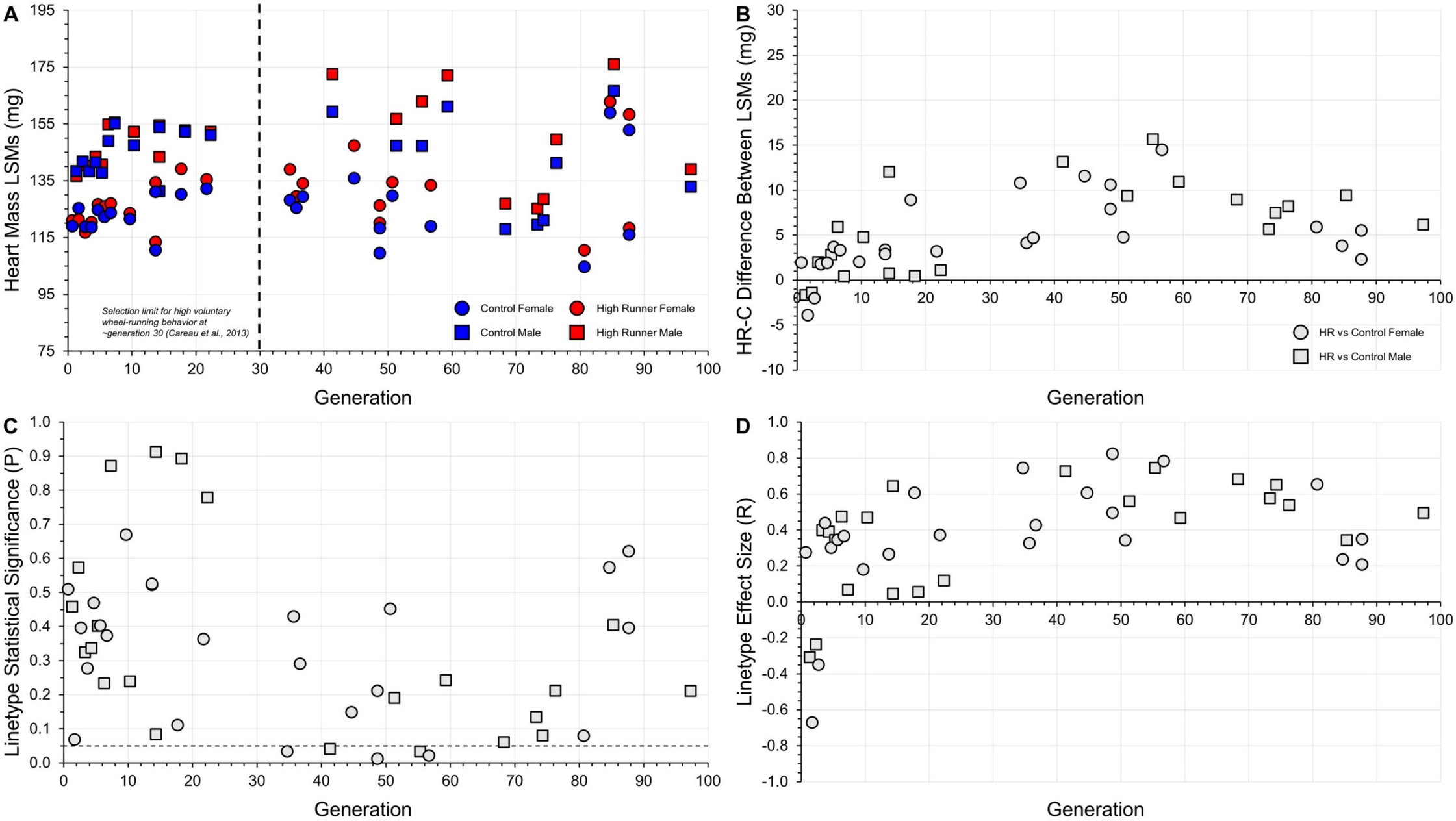
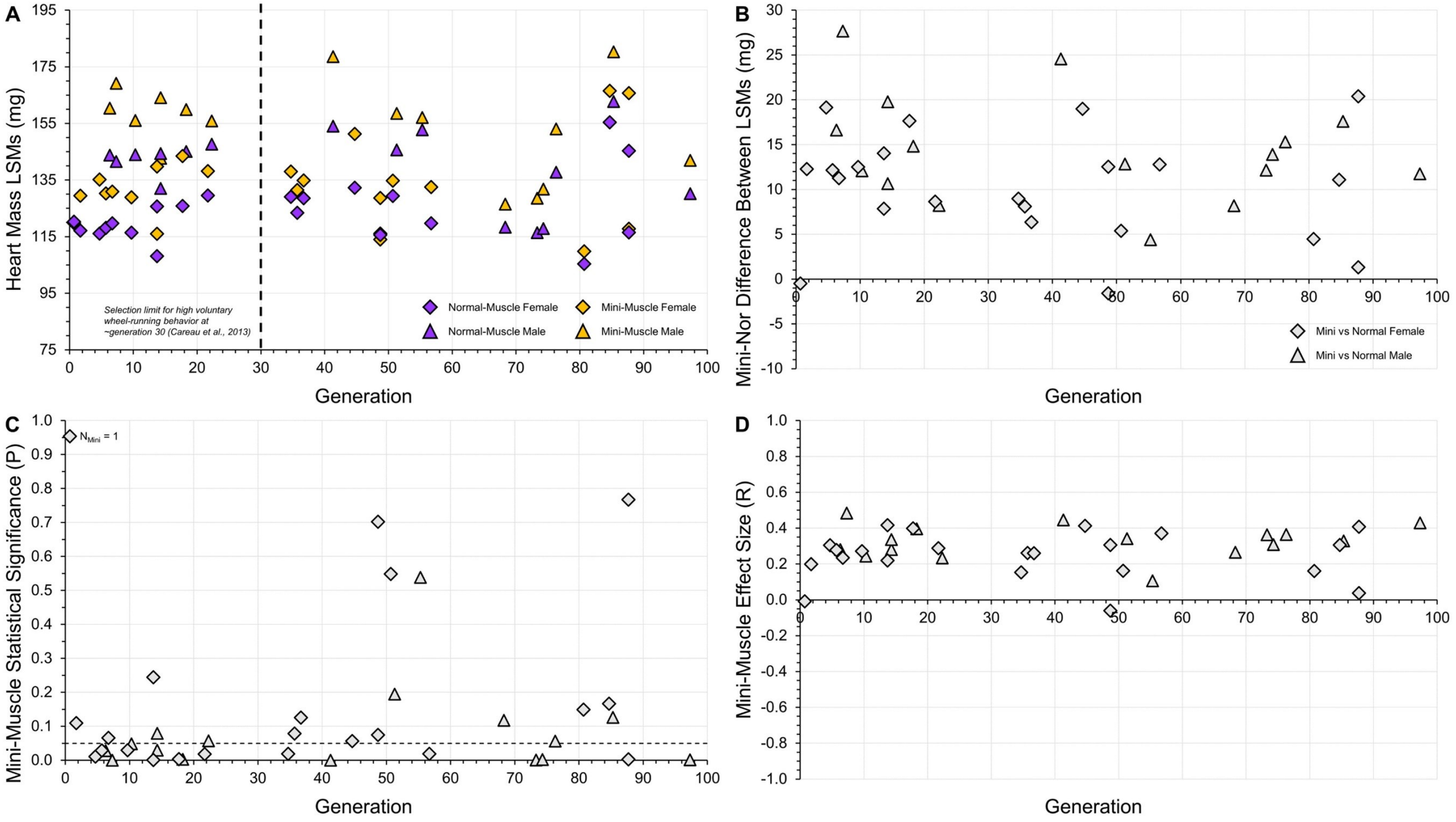

A
Heart Mass LSMs (mg)
Generation
Selection limit for high voluntary wheel-running behavior at ~generation 30 (Careau et al., 2013)
Control Female
High Runner Female
Control Male
High Runner Male
B
HR-C Difference Between LSMs (mg)
Generation
HR vs Control Female
HR vs Control Male
C
Linetype Statistical Significance (P)
Generation
D
Linetype Effect Size (R)
Generation

Figure 4. Values for heart (ventricle) mass (with body mass as a covariate) based on mini-muscle status (mini vs normal). For more information, see descriptions in Figure 2.


A
Heart Mass LSMs (mg)
Generation
Selection limit for high voluntary wheel-running behavior at ~generation 30 (Careau et al., 2013)
Normal-Muscle Female
Mini-Muscle Female
Normal-Muscle Male
Mini-Muscle Male
B
Mini-Nor Difference Between LSMs (mg)
Generation
Mini vs Normal Female
Mini vs Normal Male
C
Mini-Muscle Statistical Significance (P)
Generation
N_Mini = 1
D
Mini-Muscle Effect Size (R)
Generation

References

Agrawal, A. A. 2020. A scale-dependent framework for trade-offs, syndromes, and specialization in organismal biology. Ecology 101:e02924.

Bartolucci, A. A., M. Tendera, and G. Howard. 2011. Meta-analysis of multiple primary prevention trials of cardiovascular events using aspirin. American Journal of Cardiology 107:1796–1801.

Belke, T. W., and T. Garland Jr. 2007. A brief opportunity to run does not function as a reinforcer for mice selected for high daily wheel-running rates. J Exp Anal Behav 88:199–213.

Bennett, A. F. 2003. Experimental evolution and the Krogh Principle: Generating biological novelty for functional and genetic analyses. Physiological and Biochemical Zoology 76:1–11. The University of Chicago Press.

Bilodeau, G. M., H. Guderley, D. R. Joanisse, and T. Garland Jr. 2009. Reduction of type IIb myosin and IIB fibers in tibialis anterior muscle of mini-muscle mice from high-activity lines. J. Exp. Zool. 311A:189–198.

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò. 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365–376.

Cadney, M. D., R. L. Albuquerque, N. E. Schwartz, M. P. McNamara, A. A. Castro, M. P. Schmill, D. A. Hillis, and T. Garland Jr. 2022. Effects of early-life voluntary exercise and fructose on adult activity levels, body composition, aerobic capacity, and organ masses in mice bred for high voluntary wheel-running behavior. Journal of Developmental Origins of Health and Disease 1–12.

Cadney, M. D., L. Hiramatsu, Z. Thompson, M. Zhao, J. C. Kay, J. M. Singleton, R. L. Albuquerque, M. P. Schmill, W. Saltzman, and T. Garland Jr. 2021. Effects of early-life exposure to Western diet and voluntary exercise on adult activity levels, exercise physiology, and associated traits in selectively bred High Runner mice. Physiology & Behavior 234:113389.

Careau, V., M. E. Wolak, P. A. Carter, and T. Garland Jr. 2013. Limits to behavioral evolution: The quantitative genetics of a complex trait under directional selection. Evolution 67:3102–3119.

Castro, A. A., and T. Garland Jr. 2018. Evolution of hindlimb bone dimensions and muscle masses in house mice selectively bred for high voluntary wheel-running behavior. Journal of Morphology 279:766–779.

Castro, A. A., F. A. Karakostis, L. E. Copes, H. E. McClendon, A. P. Trivedi, N. E. Schwartz, and T. Garland Jr. 2021a. Effects of selective breeding for voluntary exercise, chronic exercise, and their interaction on muscle attachment site morphology in house mice. Journal of Anatomy 240:279–295.

Castro, A. A., A. Nguyen, S. Ahmed, T. Garland Jr., and N. C. Holt. 2024. Muscle-tendon unit properties in mice selected for high levels of voluntary running: novel physiologies, co-adaptation, trade-offs, and multiple solutions in the evolution of cursoriality. Ecological and Evolutionary Physiology In press.

Castro, A. A., H. Rabitoy, G. C. Claghorn, and T. Garland Jr. 2021b. Rapid and longer-term effects of selective breeding for voluntary exercise behavior on skeletal morphology in house mice. J Anat 238:720–742.

Claghorn, G. C., I. A. T. Fonseca, Z. Thompson, C. Barber, and T. Garland Jr. 2016. Serotonin-mediated central fatigue underlies increased endurance capacity in mice from lines selectively bred for high voluntary wheel running. Physiology & Behavior 161:145–154.

Cloyed, C. S., J. M. Grady, V. M. Savage, J. C. Uyeda, and A. I. Dell. 2021. The allometry of locomotion. Ecology 102:e03369.

Cohen, A. A., C. F. D. Coste, X.-Y. Li, S. Bourg, and S. Pavard. 2020. Are trade-offs really the key drivers of ageing and life span? Functional Ecology 34:153–166.

Cohen, J. 1988. Statistical power analysis for the behavioral sciences. 2nd ed. L. Erlbaum Associates, Hillsdale, N.J.

Copes, L., H. Schutz, E. Dlugosz, W. Acosta, M. Chappell, and T. Garland Jr. 2015. Effects of voluntary exercise on spontaneous physical activity and food consumption in mice: Results from an artificial selection experiment. Physiology & Behavior 149:86–94.

Cruden, K., K. Wilkinson, D. K. Mukaz, T. B. Plante, N. A. Zakai, D. L. Long, M. Cushman, and N. C. Olson. 2024. Soluble CD14 and incident diabetes risk: The REasons for Geographic and Racial Differences in Stroke (REGARDS) study. J Endocr Soc 8:bvae097.

Djawdan, M. 1993. Locomotor performance of bipedal and quadrupedal heteromyid rodents. Functional Ecology 7:195–202.

Garland Jr., T. 2003. Selection experiments: An under-utilized tool in biomechanics and organismal biology. Pp. 23–56 *in* Vertebrate Biomechanics and Evolution. BIOS Scientific Publishers Limited, Oxford.

Garland Jr., T., and R. L. Albuquerque. 2017. Locomotion, energetics, performance, and behavior: A mammalian perspective on lizards, and vice versa. Integr Comp Biol 57:252–266.

Garland Jr., T., M. Cadney, and R. Waterland. 2017. Early-life effects on adult physical activity: Concepts, relevance, and experimental approaches. Physiological and Biochemical Zoology 90:1–14.

Garland Jr., T., and P. A. Carter. 1994. Evolutionary physiology. Annual Review of Physiology 56:579–621.

Garland Jr., T., S. A. Kelly, J. L. Malisch, E. M. Kolb, R. M. Hannon, B. K. Keeney, S. L. Van Cleave, and K. M. Middleton. 2011a. How to run far: Multiple solutions and sex-specific responses to selective breeding for high voluntary activity levels. Proceedings of the Royal Society B: Biological Sciences 278:574–581.

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

Garland Jr., T., H. Schutz, M. Chappell, B. Keeney, T. Meek, L. Copes, W. Acosta, C. Drenowatz, R. Maciel, G. van Dijk, C. Kotz, and J. Eisenmann. 2011b. The biological control of voluntary exercise, spontaneous physical activity and daily energy expenditure in relation to obesity: Human and rodent perspectives. Journal of Experimental Biology 214:206–229.

Garland Jr., T., M. Zhao, and W. Saltzman. 2016. Hormones and the evolution of complex traits: Insights from artificial selection on behavior. Integr Comp Biol 56:207–224.

Gibbs, A. G. 1999. Laboratory selection for the comparative physiologist. Journal of Experimental Biology 202:2709–2718.

Gillooly, J. F., J. P. Gomez, and E. V. Mavrodiev. 2017. A broad-scale comparison of aerobic activity levels in vertebrates: endotherms versus ectotherms. Proc. R. Soc. B 284:20162328.

Girard, I., E. Rezende, and T. Garland Jr. 2007. Leptin levels and body composition of mice selectively bred for high voluntary locomotor activity. Physiological and Biochemical Zoology 80:568–579.

Girard, I., J. Swallow, P. Carter, P. Koteja, J. Rhodes, and T. Garland Jr. 2002. Maternal-care behavior and life-history traits in house mice (*Mus domesticus*) artificially selected for high voluntary wheel-running activity. Behavioural Processes 57:37–50.

Glass, G. 1976. Primary, secondary, and meta-analysis of research. Educational Researcher 5:3–8.

Goh, J., J. Hall, and R. Rosenthal. 2016. Mini meta-analysis of your own studies: Some arguments on why and a primer on how: Mini meta-analysis. Social and Personality Psychology Compass 10:535–549.

Guderley, H., P. Houle-Leroy, G. M. Diffee, D. M. Camp, and T. Garland Jr. 2006. Morphometry, ultrastructure, myosin isoforms, and metabolic capacities of the “mini muscles” favoured by selection for high activity in house mice. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology 144:271–282.

Guderley, H., D. R. Joanisse, S. Mokas, G. M. Bilodeau, and T. Garland Jr. 2008. Altered fibre types in gastrocnemius muscle of high wheel-running selected mice with mini-muscle phenotypes. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology 149:490–500.

Halsey, L., D. Curran-Everett, S. Vowler, and G. Drummond. 2015. The fickle P value generates irreproducible results. Nat Methods 12:179–185.

Halsey, L. G. 2019. The reign of the *p* -value is over: What alternative analyses could we employ to fill the power vacuum? Biol. Lett. 15:20190174.

Hannon, R. M., T. H. Meek, W. Acosta, R. C. Maciel, H. Schutz, and T. Garland Jr. 2011. Sex-specific heterosis in line crosses of mice selectively bred for high locomotor activity. Behav Genet 41:615–624.

Hillis, D. A., and T. Garland Jr. 2023. Multiple solutions at the genomic level in response to selective breeding for high locomotor activity. Genetics 223:iyac165.

Hillis, D. A., L. Yadgary, G. M. Weinstock, F. Pardo-Manuel de Villena, D. Pomp, A. S. Fowler, S. Xu, F. Chan, and T. Garland Jr. 2020. Genetic basis of aerobically supported voluntary exercise: results from a selection experiment with house mice. Genetics genetics.303668.2020.

Hillis, D. A., L. Yadgary, G. M. Weinstock, F. Pardo-Manuel de Villena, D. Pomp, and T. Garland Jr. 2024. Large changes in detected selection signatures after a selection limit in mice bred for voluntary wheel-running behavior. PLOS ONE. In press.

Hillman, S. S., and M. S. Hedrick. 2015. A meta-analysis of in vivo vertebrate cardiac performance: implications for cardiovascular support in the evolution of endothermy. Journal of Experimental Biology 218:1143–1150.

Hiramatsu, L., J. Kay, Z. Thompson, J. Singleton, G. Claghorn, R. Albuquerque, B. Ho, B. Ho, G. Sanchez, and T. Garland Jr. 2017. Maternal exposure to Western diet affects adult body composition and voluntary wheel running in a genotype-specific manner in mice. Physiology & Behavior 179:235–245.

Houle-Leroy, P., H. Guderley, J. G. Swallow, and T. Garland Jr. 2003. Artificial selection for high activity favors mighty mini-muscles in house mice. American Journal of Physiology-Regulatory, Integrative and Comparative Physiology 284:R433–R443.

Keeney, B. K. 2011. Behavioral, neural, and life history correlates of selective breeding for high voluntary exercise in house mice. University of California, Riverside.

Kelly, S. A., T. A. Bell, S. R. Selitsky, R. J. Buus, K. Hua, G. M. Weinstock, T. Garland Jr., F. Pardo-Manuel de Villena, and D. Pomp. 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.

Khan, R. H., J. S. Rhodes, I. A. Girard, N. E. Schwartz, and T. Garland Jr. 2024. Does behavior evolve first? Correlated responses to selection for voluntary wheel-running behavior in house mice. Ecological and Evolutionary Physiology 97:97–117.

Kolb, E. M., S. A. Kelly, K. M. Middleton, L. S. Sermsakdi, M. A. Chappell, and T. Garland Jr. 2010. Erythropoietin elevates VO2max but not voluntary wheel running in mice. Journal of Experimental Biology 213:510–519.

Kolb, E., E. Rezende, L. Holness, A. Radtke, S. Lee, A. Obenaus, and T. Garland Jr. 2013. Mice selectively bred for high voluntary wheel running have larger midbrains: Support for the mosaic model of brain evolution. Journal of Experimental Biology 216:515–523.

Konczal, M., P. Koteja, P. Orlowska-Feuer, J. Radwan, E. T. Sadowska, and W. Babik. 2016. Genomic response to selection for predatory behavior in a mammalian model of adaptive radiation. Molecular Biology and Evolution 33:2429–2440.

Koteja, P., J. Swallow, P. Carter, and T. Garland Jr. 1999. Energy cost of wheel running in house mice: implications for coadaptation of locomotion and energy budgets. Physiological and Biochemical Zoology 72:238–249.

Lakens, D. 2013. Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Front. Psychol. 4.

Langerhans, R. B. 2008. Coevolution. Pp. 32–36 *in* B. Fath, ed. Encyclopedia of Ecology (Second Edition). Elsevier, Oxford.

Lenski, R. E. 2017. Convergence and divergence in a long-term experiment with bacteria. The American Naturalist 190:S57–S68.

Malisch, J., C. Breuner, E. Kolb, H. Wada, R. Hannon, M. Chappell, K. Middleton, and T. Garland Jr. 2009. Behavioral despair and home-cage activity in mice with chronically elevated baseline corticosterone concentrations. Behavior Genetics 39:192–201.

Malisch, J., W. Saltzman, F. Gomes, E. Rezende, D. Jeske, and T. Garland Jr. 2007. Baseline and stress-induced plasma corticosterone concentrations of mice selectively bred for high voluntary wheel running. Physiological and Biochemical Zoology 80:146–156.

Marchini, M., L. M. Sparrow, M. N. Cosman, A. Dowhanik, C. B. Krueger, B. Hallgrimsson, and C. Rolian. 2014. Impacts of genetic correlation on the independent evolution of body mass and skeletal size in mammals. BMC Evolutionary Biology 14:258.

Meek, T. H., J. C. Eisenmann, and T. Garland Jr. 2010. Western diet increases wheel running in mice selectively bred for high voluntary wheel running. Int J Obes 34:960–969.

Most, P. J. van der, B. de Jong, H. K. Parmentier, and S. Verhulst. 2011. Trade-off between growth and immune function: a meta-analysis of selection experiments. Functional Ecology 25:74–80.

Mosteller, F., and R. A. Fisher. 1948. Questions and answers. The American Statistician 2:30.

Palma-Vera, S. E., H. Reyer, M. Langhammer, N. Reinsch, L. Derezanin, J. Fickel, S. Qanbari, J. M. Weitzel, S. Franzenburg, G. Hemmrich-Stanisak, and J. Schoen. 2022. Genomic characterization of the world's longest selection experiment in mouse reveals the complexity of polygenic traits. BMC Biology 20:52.

Pearson, K. 1904. Report on certain enteric fever inoculation statistics. BMJ 2:1243–1246.

Poole, D. C., and H. H. Erickson. 2011. Highly athletic terrestrial mammals: Horses and dogs. Pp. 1–37 *in* Comprehensive Physiology. John Wiley & Sons, Ltd.

Rauser, C., L. Mueller, M. Travisano, and M. Rose. 2009. Evolution of aging and late life. Experimental Evolution: Concepts, Methods, and Applications of Selection Experiments, doi: 10.1525/california/9780520247666.003.0018.

Rezende, E., F. Gomes, M. Chappell, and T. Garland Jr. 2009. Running behavior and its energy cost in mice selectively bred for high voluntary locomotor activity. Physiological and Biochemical Zoology 82:662–679.

Rezende, E. L., M. A. Chappell, F. R. Gomes, J. L. Malisch, and T. Garland Jr. 2005. Maximal metabolic rates during voluntary exercise, forced exercise, and cold exposure in house mice selectively bred for high wheel-running. Journal of Experimental Biology 208:2447–2458.

Rezende, E. L., T. Garland Jr., M. A. Chappell, J. L. Malisch, and F. R. Gomes. 2006. Maximum aerobic performance in lines of *Mus* selected for high wheel-running activity: effects of selection, oxygen availability and the mini-muscle phenotype. Journal of Experimental Biology 209:115–127.

Rhodes, J. S., S. C. Gammie, and T. Garland Jr. 2005. Neurobiology of mice selected for high voluntary wheel-running activity. Integrative and Comparative Biology 45:438–455.

Rhodes, J. S., and T. J. Kawecki. 2009. Behavior and neurobiology. Pp. 263–300 *in* T. Garland Jr. and M. R. Rose, eds. Experimental evolution: Concepts, methods, and applications of selection experiments. University of California Press.

Rice, W. R., and B. Holland. 1997. The enemies within: intergenomic conflict, interlocus contest evolution (ICE), and the intraspecific Red Queen. Behav Ecol Sociobiol 41:1–10.

Rosenthal, R., and R. L. Rosnow. 2008. Essentials of behavioral research: Methods and data analysis. 3rd ed. McGraw-Hill, Boston.

Sadowska, E. T., K. Baliga-Klimczyk, K. M. Chrząścik, and P. Koteja. 2008. Laboratory model of adaptive radiation: a selection experiment in the Bank Vole. Physiological and Biochemical Zoology 81:627–640.

Sadowska, E. T., C. Stawski, A. Rudolf, G. Dheyongera, K. M. Chrząścik, K. Baliga-Klimczyk, and P. Koteja. 2015. Evolution of basal metabolic rate in bank voles from a multidirectional selection experiment. Proceedings of the Royal Society B: Biological Sciences 282:20150025.

Schmill, M. P., Z. Thompson, D. Lee, L. Haddadin, S. Mitra, R. Ezzat, S. Shelton, P. Levin, S. Benham, K. J. Huffman, and T. Garland Jr. 2023. Hippocampal, whole midbrain, red nucleus, and ventral tegmental area volumes are increased by selective breeding for high voluntary wheel-running behavior. Genes, Brain and Behavior 98:245–263.

Schwartz, N. E., M. P. McNamara, J. M. Orozco, J. O. Rashid, A. P. Thai, and T. Garland Jr. 2023. Selective breeding for high voluntary exercise in mice increases maximal ($\dot{V}O_{2,max}$) but not basal metabolic rate. Journal of Experimental Biology 226:1–10.

Serdar, C. C., M. Cihan, D. Yücel, and M. A. Serdar. 2021. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. Biochem Med (Zagreb) 31:010502.

Sinervo, B., and E. Svensson. 2002. Correlational selection and the evolution of genomic architecture. Heredity 89:329–338.

Storz, J. F., J. T. Bridgham, S. A. Kelly, and T. Garland Jr. 2015. Genetic approaches in comparative and evolutionary physiology. American Journal of Physiology-Regulatory, Integrative and Comparative Physiology 309:R197–R214.

Sullivan, G. M., and R. Feinn. 2012. Using effect size—or why the p-value is not enough. J Grad Med Educ 4:279–282.

Swallow, J. G., P. A. Carter, and T. Garland Jr. 1998. Artificial selection for increased wheel-running behavior in house mice. Behavior Genetics 28:227–237.

Swallow, J. G., J. P. Hayes, P. Koteja, and T. Garland Jr. 2009. Selection experiments and experimental evolution of performance and physiology. Pp. 301–351 *in* T. Garland Jr. and M. R. Rose, eds. Experimental evolution: Concepts, methods, and applications of selection experiments. University of California Press.

Swallow, J., P. Koteja, P. Carter, and T. Garland Jr. 1999. Artificial selection for increased wheel-running activity in house mice results in decreased body mass at maturity. Journal of Experimental Biology 202:2513–2520.

Swallow, J., P. Koteja, P. Carter, and T. Garland Jr. 2001. Food consumption and body composition in mice selected for high wheel-running activity. Journal of Comparative Physiology B: Biochemical, Systemic, and Environmental Physiology 171:651–659.

Talmadge, R. J., W. Acosta, and T. Garland Jr. 2014. Myosin heavy chain isoform expression in adult and juvenile mini-muscle mice bred for high-voluntary wheel running. Mechanisms of Development 134:16–30.

Vaanholt, L., P. Meerlo, T. Garland Jr., G. Visser, and G. van Dijk. 2007. Plasma adiponectin is increased in mice selectively bred for high wheel-running activity, but not by wheel running *per sé*. Hormone and Metabolic Research 39:377–383.

van der Most, P. J., B. de Jong, H. K. Parmentier, and S. Verhulst. 2011. Trade-off between growth and immune function: a meta-analysis of selection experiments. Functional Ecology 25:74–80.

Van Valen, L. 1973. A new evolutionary law. Evolutionary Theory 1:1–30.

Wahlsten, D. 1990. Insensitivity of the analysis of variance to heredity-environment interaction. Behavioral and Brain Sciences 13:109–120.

Wallace, I. J., and T. Garland Jr. 2016. Mobility as an emergent property of biological organization: Insights from experimental evolution. Evolutionary Anthropology 25:98–104.

Wasserstein, R. L., and N. Lazar. 2016. The ASA's statement on *p* -values: Context, process, and purpose. The American Statistician 70:129–133.

Weber, K. E. 1990. Selection on wing allometry in Drosophila melanogaster. Genetics 126:975–989.

White, C. R., P. Cassey, and T. M. Blackburn. 2007. Allometric exponents do not support a universal metabolic allometry. Ecology 88:315–323.

Wone, B. W. M., P. Madsen, E. R. Donovan, M. K. Labocha, M. W. Sears, C. J. Downs, D. A. Sorensen, and J. P. Hayes. 2015. A strong response to selection on mass-

independent maximal metabolic rate without a correlated response in basal metabolic rate. Heredity (Edinb) 114:419–427.