Large changes in detected selection signatures after a selection limit in mice bred for voluntary wheel-running behavior

Short title: Evolution following reaching a selection limit

David A. Hillis[1*], Liran Yadgary[2,#a], George M. Weinstock[3], Fernando Pardo-Manuel de Villena[2], Daniel Pomp[2], and Theodore Garland, Jr.[4]

[1] Genetics, Genomics, and Bioinformatics Graduate Program, University of California, Riverside, California 92521

[2] Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

[3] The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032 and Department of Genetics and Genome Science, University of Connecticut Health Center, Farmington, CT 06030

[4] Department of Evolution, Ecology, and Organismal Biology, University of California, Riverside, California 92521

[#a] Current Address: Hazera Seeds Ltd. in Israel, Berurim M.P Shikmim, Israel 7983700

* Corresponding author

E-mail: davidhillis@ucsb.edu

Abstract

In various organisms, sequencing of selectively bred lines at apparent selection limits has demonstrated that genetic variation can remain at many loci, implying that evolution at the genetic level may continue even if the population mean phenotype remains constant. We compared selection signatures at generations 22 and 61 of the "High Runner" mouse experiment, which includes 4 replicate lines bred for voluntary wheel-running behavior (HR) and 4 non-selected control (C) lines. Previously, we reported multiple regions of differentiation between the HR and C lines, based on whole-genome sequence data for 10 mice from each line at generation 61, which was >31 generations after selection limits had been reached in all HR lines. Here, we analyzed pooled sequencing data from ~20 mice for each of the 8 lines at generation 22, around when HR lines were reaching limits. Differentiation analyses of allele frequencies at ~4.4 million SNP loci used the regularized T-test and detected 258 differentiated regions with FDR = 0.01. Comparable analyses involving pooling generation 61 individual mouse genotypes into allele frequencies by line produced only 11 such regions, with almost no overlap among the largest and most statistically significant peaks between the two generations. These results implicate a sort of "genetic churn" that continues at loci relevant for running. Simulations indicate that loss of statistical power due to random genetic drift and sampling error are insufficient to explain the differences in selection signatures. The 13 differentiated regions at generation 22 with strict culling measures include 79 genes related to a wide variety of functions. Gene ontology identified pathways related to olfaction and vomeronasal pathways as being overrepresented, consistent with generation 61 analyses, despite those specific regions differing between

54 generations. Genes *Dspp* and *Rbm24* are also identified as potentially explaining

55 known bone and skeletal muscle differences, respectively, between the linetypes.

56
57
58
59

Introduction

Although evolution can result in organisms with spectacular capabilities or able to survive in exceptionally inhospitable environments, all adaptations are bound within certain limits. These limits are commonly observed in laboratory and agricultural selection experiments (Dobzhansky and Spassky 1969; Al-Murrani and Roberts 1974; Careau et al. 2013; Schlötterer et al. 2015; Lillie et al. 2019). Among various possible causes of selection limits (Al-Murrani and Roberts 1974; Falconer 1989; Douhard et al. 2021), the simplest explanation is the loss of genetic variation, such that narrow-sense heritability declines to zero (e.g., see Brown and Bell 1961). However, selection experiments have frequently found that genetic variation remains after reaching a selection limit (e.g., Lerner and Dempster 1951; Roberts 1966; Dobzhansky and Spassky 1969; Bult and Lynch 2000; Burke et al. 2010; Careau et al. 2013; Lillie et al. 2019; Hillis et al. 2020). Even for alleles favored by selection, fixation is far from guaranteed (Burke et al. 2010; Schlötterer et al. 2015; Stephan 2016; Hillis et al. 2020).

One selection experiment that has continued selection long after reaching a limit is the High Runner (HR) mouse experiment, which started in 1993 with the purchase of 224 outbred ICR mice from Harlan Sprague Dawley (Swallow, Carter, et al. 1998). These were randomly bred for two generations, then split into ten breeding pairs to found each of eight closed lines. Four of these lines were designated to serve as non-selected control lines, while the other four were selected based on voluntary wheel running. In selected lines, all mice are given access to wheels for 6 days and the male and female of each family with the highest running on days 5 and 6 would be used as breeders (no sib-mating). After about 22 generations of selection, three of the four HR

lines (with the fourth line following suit a few generations later) had plateaued in their running at approximately 2.5 to 3 times as many revolutions as the controls (Careau et al. 2013). Recently, the experiment has reached its 100th generation since selection began and, with exception of some generations when the experiment moved from Wisconsin to California (generations 32 to 35) and during Covid-19 lockdowns (generations 91 to 98), selection has continued nearly uninterrupted in the interim. Whether selection interruption following the move to California resulted in statistically significant changes to running behavior has not yet been analyzed.

Numerous physiological and morphological differences between the HR and control lines have been documented (Rhodes et al. 2005; Swallow et al. 2009; Garland, Jr., Schutz, et al. 2011; Wallace and Garland, Jr. 2016; Khan et al. 2024). These include traits associated with motivation to run, such as changes in dopamine (Rhodes et al. 2001; Mathes et al. 2010), serotonin (Waters et al. 2013), and endocannabinoid signaling (Thompson et al. 2017), as well as changes in brain size and structure (Kolb, Rezende, et al. 2013). Additionally, changes associated with ability to run have been found, including endurance capacity (Meek et al. 2009), maximal aerobic capacity (VO_{2Max}) (e.g., Swallow, Garland, Jr., et al. 1998; Kolb et al. 2010; Dlugosz et al. 2013; Hiramatsu et al. 2017; Cadney et al. 2021), heart size (Kolb et al. 2010; Kolb, Kelly, et al. 2013; Kelly et al. 2017), skeletal muscle physiology (Dumke et al. 2001; Syme et al. 2005; Guderley et al. 2008; Castro et al. 2022), and bone morphology (Garland, Jr. and Freeman 2005; Kelly et al. 2006; Middleton et al. 2008; Middleton et al. 2010; Wallace et al. 2010; Wallace et al. 2012; Castro and Garland, Jr. 2018; Copes et al. 2018; Schwartz et al. 2018).

Previously, whole-genome differentiation analyses using individual mouse data from 10 males from each of the eight lines at generation 61 identified at least 13 genomic regions differentiated between the control and HR lines (Hillis et al. 2020; Hillis and Garland, Jr. 2022). Within these regions were genes associated with development of the brain, heart, bones, and limbs, in addition to reward pathways, and even the vomeronasal system (see also Nguyen et al. 2020). Dropping individual lines from analyses revealed new potential signatures of selection and demonstrated that the HR lines have evolved in different ways at the genomic level ("multiple solutions" Garland, Jr., Kelly, et al. 2011) that increase wheel-running behavior (Hillis and Garland, Jr. 2022). Despite being ~30-35 generations past the selection limit, a great deal of genetic diversity remained in all 8 lines including many regions identified as differentiated between the HR lines and controls.

With the selection limit achieved near generation 22, one might expect many if not most biologically relevant SNPs to already be differentiated by that generation. Thus, with respect to the ability to detect selection signatures, little advantage would be gained from allowing ~30-35 generations to pass before testing for allelic differentiation between the HR and control lines. Furthermore, simulations performed by Baldwin-Brown et al. (2014) demonstrate that increasing the number of generations could reduce power to detect some loci under selection, which they attributed to noise created by random genetic drift. Reasonably, one might expect that drift over enough generations may cause control lines to diverge from each other in allele frequencies, such that selection signatures are obscured in statistical tests that compare replicate sets of selected and control lines. For example, if some control lines become fixed for

129 one allele and the remaining control lines become fixed for another, then, even if all HR
130 lines were fixed for the same allele favored by selection, statistically significant
131 differentiation would be difficult to detect. Therefore, analyses of a generation close to
132 when a selection limit is first reached would be optimal for tests of genetic
133 differentiation.

134 In the present study, we analyze pooled sequence data from each of the four HR
135 lines and four control lines at generation 22. Although these analyses identify many
136 regions containing genes associated with systems known to be phenotypically
137 differentiated between the HR and control linetypes, they largely differ from those
138 previously identified with the generation 61 individual mouse sequence data (Hillis et al.
139 2020). Furthermore, the number of differentiated regions detected at generation 22 are
140 more than 20-fold greater than those detected with generation 61 data (treated as
141 pooled data).

142 We first discuss possible methodological causes of these differences (e.g.,
143 pooled vs individual mouse data) and find them lacking. We therefore develop a simple
144 simulation model, with leptokurtic distribution of locus effect sizes, to test the possibility
145 that a hypothetical physiological constraint on wheel running could contribute to the
146 differences between generations 22 and 61 selection signatures. Ignoring locus effect
147 size, results demonstrate that such constraints can contribute to a reduction in power
148 and increased variability in the detected response to selection in generations after the
149 selection limit. However, the magnitude of these effects appears insufficient to explain
150 the differences observed between generations 22 and 61 in the real data. In addition,
151 effect size was an important determinant of the ability to detect selection signatures in

the simulations, including a more than 2-fold increase in power to detect loci with large effect size at generation 22 as compared to generation 61. Thus, with strict culling procedures, we suspect that many of the selection signatures detected at both generations are likely to represent loci with relatively large effects on wheel running. The regions detected at generation 22 include genes related to olfactory/vomeronasal systems, which are also identified at generation 61 (Hillis et al. 2020; Nguyen et al. 2020; Hillis and Garland, Jr. 2022).

Materials and methods

High runner mouse model

As described previously (Careau et al. 2013; Swallow, Carter, et al. 1998), 112 males and 112 females of the outbred Hsd:ICR strain were purchased from Harlan Sprague Dawley in 1993 and designated as generation -2. Mice would be randomly bred for 2 generations (-2 and -1) with 2-3 generation -1 mice from each family randomly chosen to contribute to 1 of 8 different closed lines. Four of these lines were randomly picked to be "High Runner" (HR) lines, in which mice would be selected for breeding based on voluntary wheel running. The remaining 4 lines were used as Control (C) lines, without any selection. Generation 0 was the first generation where HR lines were paired based on running levels (10 males and 10 females for each line) with generation 1 the first product of selection.

Wheel running measurements were collected by giving mice at approximately 6-8 weeks of age, access to wheels for six days. The amount of running (total revolutions) on days 5 and 6 was used as the selection criterion. Both days 5 and 6 are used for repeatability in running behavior and robustness against bad data for a single day (Careau et al. 2013). For the HR lines, the highest-running male and female from within each of 10 families were chosen as breeders (within-family selection). For the non-selected C lines, one male and one female from each of 10 families were chosen as breeders, independent of wheel running measurements. Sib-mating was disallowed in all lines (Swallow, Carter, et al. 1998).

Genome sequencing and allele frequency determination

Roughly 10 male and 10 female mice were taken from each line at generation 22 (Khan et al. 2024). Mice were decapitated without anesthesia because blood was being taken for a study of hormone levels (corticosterone) that can respond rapidly to additional handling or anesthesia. Subsequently, their DNA was extracted from tail tips and then pooled for determination of allele frequency for each line. This pooled DNA was sequenced with paired end pooled sequencing with Illumina HiSeq 2500 sequences were trimmed and aligned to the GRCm38/mm10 mouse genome assembly. Generation 22 used trimmomatic v0.39 for trimming, BWA v0.7.17 for alignment, Samtools v1.14 for sorting and indexing, picard v2.26.11 for marking duplicates, and GATK v4.1.8.1 for calling SNPs. SNPs were filtered to keep those with read quality ("RQ") ≥ 20, DP ≥ 10, were missing either quality score, or missing the allele frequency all together, or had MAF > 0.0126. Allele frequencies ("AF") were determined for generation 22 by taking the read depth of the alternate nucleotide allele (i.e., allele differing from the GRCm38/mm10 alignment) and dividing by the read depth for the locus. After all quality control methods were implemented, 4,446,523 loci remained for generation 22.

The generation 61 data were taken from Hillis et al. (2020). 80 male mice (10 from each line) were subject to whole genome sequencing and reads were trimmed and aligned to the GRCm38/mm10 mouse genome assembly as described in Didion et al. (2016). This generated an average read depth of 12X per mouse. SNPs were filtered to keep those with genotype quality ("GQ") > 5, read depth ("DP") > 3, minimum allele frequency ("MAF") > 0.0126 for all samples, and Mapping Quality ("MQ") > 30. One of

the 80 mice was excluded due to likely contamination (as in Xu and Garland 2017), leaving 79 for the following analyses. SNPs not found to be present in at least two of the 79 mice were also removed from analysis. After all quality control methods were implemented, 5,932,148 loci remained for analyses. To allow comparison with the pooled sequencing data from generation 22, we calculated allele frequencies as the number of alternative alleles divided by 2 times the number of mice (i.e., 20 or 18 for HR3).

Statistical analyses

For generations 22 and 61 we used an arcsine-squared transformation (Ahrens et al. 1990) of the AF. Analyses were conducted on both generations using a traditional T-test, regularized T-test (RegT)(Baldwin-Brown et al. 2014, see also Baldi and Long 2001), and a variant of the regularized T-test which uses a sliding window to calculate $\bar{v}$ (WRT test) (S1 File). The regularized T-test was based on a Bayesian method meant to minimize the type-I errors caused by sampling error with small sample sizes (Baldi and Long 2001; Baldwin-Brown et al. 2014), such as the 8 total lines in the HR mouse selection experiment. We performed these tests and determined the permutation-based false discovery rate (FDR) for each method (see below). For comparison, we also performed the RegT and WRT tests on loci found in both generation 22 and 61 (from pooling individual mouse genotypes) data sets along with the FDR. Since standard T-tests do not require whole genome or region variances of other loci, the p-values of loci shared between the two generations could simply be extracted from the complete original analyses.

Permutation-based false discovery rate

To determine relative power of generation 22 allele frequencies with arcsine-square transformation using T-test, regularized T-test, and WRT test, we attempted to calculate a critical threshold by estimating the FDR of 10% (Benjamini and Hochberg 1995; Xie et al. 2005). However, after calculating p-values for complete permutations of the different lines within linetype to better understand the null distribution, we concluded that this estimated FDR was underestimating the true false discovery rate. Therefore, using these same permutations, we calculated the FDR directly.

Direct calculations of FDR were performed by calculating FDR for each locus of the unpermuted data whose p-value was below 0.01 in accordance with the equation:

$$FDR = \frac{n\ False\ Positives}{n\ rejected\ Null\ Hypotheses}$$

This was implemented for each locus with:

$$FDR = \frac{\frac{n\ permuted\ loci\ significant\ at\ p}{35}}{n\ unpermuted\ loci\ significant\ at\ p}$$

Loci with nominal p-value < 0.05 were ordered by FDR score, the p-value was identified for the locus with the largest FDR below 0.01, and any p-values less than or equal to the p-value for this locus was treated as significant. The SNPs with FDR = 0.01 were then further grouped into “significant regions” by grouping any loci within 1mbp of another and separating groups whose closest SNPs are further than 1mbp.

Divergence over time

To test for a difference in the number of loci showing a significant change in allele frequency between the HR and C lines from generations 22 to 61, we first conducted a paired T-test for the 4 C lines and separately for the 4 HR lines. These tests were based on eight data points for each linetype, i.e., the mean allele frequency for each line at a given locus at generation 22 and 61. The T-score for the C T-test was then subtracted from the T-score for the HR T-test, and the absolute value was taken. This was repeated for each locus, producing values for approximately 2 million loci (excluding where either the C or HR T-test failed for numerical reasons). These analyses were then repeated with all 35 permutations (as described above) to estimate the null distribution of the score based on ~2,000,000 * 35 = ~75,000,000 values. These scores were ordered to identify the 5^{th} percentile threshold for comparison with the distribution of the unpermuted results.

"Strict" culling for biological and AF change analyses

Rather than attempt to focus on the genes of more than 100 regions for each of the different statistical tests, analyses of biological significance and comparisons of change in allele frequencies between generations 22 and 61 were done using a subset of the regions identified by FDR. WRT and regularized T-test first culled by removing regions containing only one significant locus, then culled such that only regions containing at least 20 significant loci or the lowest p-value among loci in the region was below 1.00E-04. Regions associated with the T-tests were culled in a similar manner as the WRT and RegT test, except the p-value cutoff used was 1.00E-06 due to naturally lower p-

values. These culling methods should also serve to reduce the influence of sampling error, as it would be increasingly unlikely for sampling error to simultaneously underestimate among-line variance across multiple linked SNPs and lines. We will refer to these additional culling methods below as “strict” culling.

Comparison of selection signatures in generations 22 and 61

Changes in allele frequencies from generation 22 to 61 were analyzed for each region identified by strict culling for generations 22 and 61. For regions significant at generation 22, each region and its included SNPs with nominal $p<0.05$ at generation 22 were matched with SNPs at generation 61. The allele frequencies of these SNPs were averaged for each line and generation and line graphs created (one for each line) with generation 22 AF on the left and generation 61 AF on the right. This was then repeated for regions significant at generation 61, except each region and its included SNPs with nominal $p<0.05$ at generation 61 were matched with SNPs at generation 22.

Simulations to compare presumptive statistical power across generations

The available data from the two generations differ in multiple ways that might affect cross-generation comparisons of selection signatures. Each generation, each line is reduced to ~20 individuals when ~10 breeding pairs are formed. An ideal "sample" from a given generation would include all 20 of those breeding individuals. Instead, our sample from generation 22 was of ~10 males and 10 females per line that were sampled at random at the time of weaning (i.e., they were not the 20 breeding parents). In contrast, the mice from generation 61 were a semi-random sample of 10 males from

298 each line (except nine from HR3 and one female that was unintentionally used from
299 another line) (Hillis et al. 2020).

300 For a pooled DNA sample, as for generation 22, a further ideal condition is for
301 the sample of DNA from each mouse to be of equal volume and concentration through
302 the extraction and pipetting steps prior to pooling. This would then result in each
303 mouse's alleles being represented in equal quantities in the pooled sequencing sample.

304 The next source of error is read depth, which is effectively a random sampling of
305 alleles from the pooled sample. Our generation 22 samples were read at an average
306 depth of 24X. Thus, the frequency of alternative nucleotide alleles for a given SNP
307 locus was calculated by counting the number of alternative alleles, which was taken as
308 anything other than the reference. Thus, not all of the 40 alleles (as one of two
309 possibilities) contributed by the 20 mice could have been identified with a read depth of
310 24X, which acts as 24 samples taken with replacement.

311 The generation 61 data are from individual sequencing of 10 mice per line at an
312 average read depth of 12X, with those results then used to predict the genotype for
313 each SNP and mouse (Hillis et al. 2020). This should allow for the representation of
314 nearly all alleles (N = 2 alleles x 10 mice). Originally, those data were analyzed as such
315 via mixed models to detect selection signatures (Hillis et al. 2020). Here, to allow
316 comparison with the pooled sequencing data from generation 22, we calculated allele
317 frequencies as the number of alternative alleles divided by 2 times the number of mice
318 (i.e., 20 or 18 for HR3), which should incorporate 19-20 unique alleles in equal
319 proportion. Given that the data available from the two generations differ in multiple
320 ways, we used simulations in an attempt to assess how this might affect our results.

For generation 22, simulations to elucidate possible sampling errors were performed such that alleles for 20 mice were sampled using a random binomial distribution assuming population allele frequencies of (0.05, 0.10, 0.15, …,0.90, and 0.95). Then an allele depth was randomly sampled from the actual quality data for the SNPs used in the generation 22 analyses and alleles were sampled from these simulated 20 mice (with replace) equal to this read depth. The allele frequency was then calculated as the number of alternative alleles (1) divided by the total read depth. This generated a distribution of allele frequencies given a particular starting AF for the population and was repeated 100,000 times for each starting population AF.

For generation 61, simulations were performed such that alleles for 10 mice were sampled using a random binomial distribution assuming population allele frequencies of (0.05, 0.10, 0.15, …,0.90, and 0.95). Then for each simulated mouse's genotype, a genotype quality was randomly sampled from the actual quality data for the SNPs used in the generation 61 analyses. If the simulated genotype for the mouse was heterozygous, then the genotype quality would be used to generate a 0 or 1 with the probability of a 1 equaling that of the probability of a genotyping error. If a 1 was generated (thus an error occurred) the second allele for the mouse was replace with a copy of the first allele of the mouse. The allele frequency was then calculated as the number of alternative alleles (i.e., 1) for all ten mice divided by the total alleles (i.e., 20). This generated a distribution of allele frequencies given a particular starting AF for the population and was repeated 100,000 times for each starting population AF.

Power analyses were then done by sampling four AF values from the simulated AF values from an actual population AF of 0.4 for one linetype. Likewise, four AF

values were sampled from the simulated AF values from an actual population AF of 0.6 for the other linetype. Sampled allele frequencies were transformed using an arcsine-squared transformation. A T-test (assuming unequal variance) was then conducted comparing these 8 sampled AF values. Note that this could not be done for RegT and WRT tests because it would require simulations of regional or genome-wide variance structure. These sampling and T-tests were repeated 10,000 times.

Simulations comparing power with and without a biological constraint

We used simulations to begin to address whether a biological constraint on a trait under selection (e.g., wheel running) might affect (1) the ability to detect selection signatures at generations before (e.g., generation 22) versus long after (e.g., generation 61) selection limits were reached, (2) the consistency of those signatures across generations, and (3) the rate at which loci with different allelic effect sizes respond to selection. Our rationale for using a constraint model is explained in the Discussion. As a heuristic, some of the parameters in these simulations were chosen to approximate values observed in the selection experiment and help build a model of architecture for wheel running in the HR and control mice (Sella and Barton 2019).

Running levels were calculated based on the general equation:

$$y = \mu + v_g + v_e$$

Where y is equal to the phenotype (wheel revolutions/day) of an individual mouse; μ is the "base" mean number of revolutions (held constant at the starting value set at generation 0); v_g is the variance contributed by genetic variation; and v_e is the variance contributed by environmental effects.

As a regression model, this equation is:

$$y = \mu + \beta_1 X_1 + \beta_2 X_2$$

where the genetic variance is represented by $\beta_1 X_1$ and the environmental variance is represented by $\beta_1 X_2$. X_1 represents the summed effect on wheel running of all alleles carried by the individual, where, to simulate a leptokurtic distribution (Barton and Turelli 1989; Reeve 2000; Reeve and Fairbairn 2001), these alleles are coded as having variable allelic affects (specifically, ±0.4, ±0.8, ±1.6, ±3.2, ... ±204.8) at frequencies inversely proportional to their effect size (specifically, 720 loci with effect ±0.4, 480 loci with effect ±0.8, … 8 loci with effect ±204.8) for a total of 2,096 loci, which approximates the number of haplotype blocks observed across all eight lines (Hillis et al. 2020). X_2 provides the random element of the environmental variance and is determined by randomly sampling from a normal distribution with mean = 0 and SD = 1.

The equation we applied for these simulations is:

$$y = 4{,}570 + 1.3X_1 + 2{,}100X_2$$

The values for β_1 (1.3), β_2 (2,100) and for the number of loci were determined in conjunction with one another to approximate realistic (in no particular order) (1) heritability of wheel running at the base generation being about 0.32 (Careau et al. 2013), (2) within-line coefficients of variation as being about 0.57 (Swallow, Carter, et al.

1998), and (3) realistic response to selection in the HR lines (i.e., achieving ~16,000 revolutions around generation 22)(Careau et al. 2013).

However, this equation does not adequately simulate seasonal variation (see Appendix S5 in Careau et al. 2013), so we applied an additional modifier:

$$y = S * (4{,}570 + 1.3X_1 + 2{,}100X_2)$$

S is a constant that alternates cycles between 0.769 (summer), 1 (winter and fall), and 1.3 (winter). As generation time for the first 61 generations was consistently around 3 months, these constants can alternate with each generation. The mean of 4,570 (revolutions/day) was picked to approximate the empirically determined starting running levels at generation 0 (Swallow, Carter, et al. 1998).

Any running level calculated as below 100 was set to 100, which is approximately the lowest amount of running ever observed. The maximum wheel-running for unconstrained simulations was 50,000 revolutions, which is nearly twice as high as has ever been observed in actual measurements from the selection experiment (Rhodes et al. 2003; Careau et al. 2013). In practice, the highest running level produced by the unconstrained simulations was 38,875 (of 24,400 total individuals simulated over 61 generations for the HR lines).

For the starting population of any given line, two alleles were first assigned to each of the 2,096 independently segregating starting loci for 20 mice (based on the actual selection procedures: Swallow, Carter, et al. 1998) using a random binomial distribution with p = 0.5. For control lines, mice were paired, and alleles sampled from

each of the pair to produce two male and two female offspring (to match the number of mice that are typically retained and wheel-tested in the selection experiment). The first of each sex for each family was then chosen to contribute to the next generation, which is functionally equivalent to the selection experiment, where breeders are chosen a random within family and sex for control lines. For HR lines, alleles were sampled from the parents for each of five males and five females (typical litter size is 10). Running distances were then calculated for all offspring, and the male and female with the highest running levels within each family were selected to breed for the subsequent generation (again, based on the actual selection procedure, which uses within-family selection). For both linetypes, siblings were barred from pairing (following the selection experiment). Simulations were run for 61 generations and alleles for all breeding pairs were saved at generation 0 and every 5 generations through 60, as well as generations 22 and 61. This was then repeated for 4 control lines and 4 HR lines.

We modeled the constraint on wheel running as a trait that itself can evolve. To obtain a realistic value for the constraint, we applied the same principles as for the wheel-running equation:

$$C = S * (10{,}000 + 1.0X_{C1} + 1{,}750X_{C2})$$

C is the constraint to be applied to the mouse's wheel running. S is the same seasonal multiplier used in the wheel-running equation, without which, higher running levels in winter become truncated. X_{C1} represents the genetic component of the constraint determined by (arbitrarily) 100 loci with effect sizes of ±1 (N=48), ±4 (N=24), ±11 (N=12), ±36 (N=8), ±101 (N=5), and ±306 (N=3). X_{C1} represents the environmental

component, determined by sampling from a normal distribution with mean = 0 and SD = 1 (similarly to wheel running). These values result in a narrow-sense heritability of ~0.2. Despite targeting a wheel-running constraint of about 16,000 revolutions in the HR lines, the base constraint value is set to 10,000 because the alleles that increase constraint are favored by selection in the HR lines. Thus, a lower base value is needed for HR lines to stop responding to selection at about 16,000 revolutions. For the constrained simulation, if a mouse ran more than its determined constraint then its revolutions were treated as equal to the constraint itself before picking the breeders for the next generation.

These simulations were repeated 100 times (with 4 HR lines and 4 control lines in each simulation) assuming no constraint and 100 times with the constraint (see S2 File, for parameters). T-tests assuming unequal variance between the 4 control lines and the 4 HR lines were performed at each of these “saved” generations (0, 5, 10, etc.) for the allele frequencies at each locus, with an arcsine-squared transform (Ahrens et al. 1990). Power was then calculated for each simulation at each saved generation by dividing the number of loci with $p \leq 0.05$ by the total number of loci (N = 2,096). Power was also calculated separating loci by effect size (see below).

Standardized selection differentials were calculated following Careau et al. (2013), by subtracting from the mean running for each sex and family the running level of the bred individual from that litter and dividing the difference by the standard deviation of the sex for that litter. Relative power under the constrained and unconstrained models was calculated using unpaired T-tests (unequal variance) on the previously described power calculations for each simulation and for each saved

458 generation. Relative power across generations was also calculated using unpaired T-
459 tests (unequal variance), separately for constrained and unconstrained simulations.
460 Relative consistency in detected selection signatures was calculated by first identifying
461 the specific significant loci (at a nominal $\alpha = 0.05$) at generations 22 and 61 in each
462 simulation. Then, the percentage of loci found significant at generation 22 that
463 remained significant at generation 61 was calculated. Unpaired T-tests (unequal
464 variance) were performed comparing these percentages for the constrained simulations
465 versus the unconstrained simulations. Lastly, ability to detect loci with different effect
466 sizes was compared using a T-test (unequal variance) of the number of significant loci
467 ($p \leq 0.05$) identified for each effect size and each simulation for generation 22
468 constrained vs unconstrained models, generation 61 constrained vs unconstrained
469 models, constrained generation 22 vs generation 61, and unconstrained generation 22
470 vs generation 61. For all graphs and estimates that required the calculation of a mean
471 value, missing values were excluded from the calculations. For example, if a p-value
472 could not be calculated for a given locus due to fixation across all lines for the same
473 allele, then this locus would be excluded from the power analyses.
474 Analyses were performed again implementing possible sampling error calculated
475 by the simulations to compare statistical power, as described in the previous section
476 "Simulations to Compare Presumptive Statistical Power Across Generations". This was
477 implemented by taking the actual allele frequency for each line at generations 22 and 61
478 in the simulations using the constraint model. These allele frequencies were then
479 replaced with an allele frequency sampled from the results (rounded to the nearest
480 0.05) of the population allele frequency of the sampling error simulations (i.e., 0.05,

0.10, 0.15... 0.95). For example, if the allele frequency for a given line at generation 22 (constraint model) was 0.25, then this 0.25 would be replaced by a randomly sampled estimated allele frequency from the sampling error simulations (generation 22) where 0.25 was the actual population allele frequency. Generation 61 allele frequencies were similarly replaced using the results of the generation 61 sampling error simulations.

Ethics statement

The selection experiment has been carried out in strict accordance with the approval from the Institutional Animal Care and Use Committee (IACUC) at two different institutions and under multiple protocol number. All experiments have been conducted to minimize distress to the animals. Any injuries or illness were treated in accordance with veterinarian recommendations. The present manuscript uses only published sequence data and new sequence data from historical tissue samples.

Results

Basic characteristics of genetic variation

The number of variable loci used in the present study includes 4,446,523 for generation 22 and 5,932,148 for generation 61. Generation 61 data had an average read depth of 12X per mouse for 10 mice in each of the 8 lines, producing an average read depth of over 100 per line for detection of many more variable SNPs in each line. The overlap of base positions between generations 22 and 61 was 2,045,546 SNPs. As expected, minor allele frequency (MAF), generally decreases for both HR and C lines between generations 22 and 61 (Fig 1). MAF values for HR and C lines are generally similar at generation 22; however, these diverge for many regions by generation 61.

Fig 1. Average minor allele frequencies. Average minor allele frequencies for generation 22 control lines, generation 22 HR lines, generation 61 control lines, and generation 61 HR lines by chromosome (numbered on the right). Regions identified as differentiated at generation 22 are indicated with an orange line above each chromosome's graph (regions smaller than 50 kbp are omitted). Regions identified as consistently differentiated at generation 61 (Hillis et al., 2020) are indicated similarly with a green line.

Differentiated SNPs and chromosomal regions

For analyses containing all generation 22 loci (N = 4,446,523), WRT identified 1,184 differentiated loci based on 0.01 FDR (Table 1). These loci fall into 258 unique regions (separated by at least 1 million base pairs). At generation 61, 1,449 loci were identified

517 as differentiated based on 0.01 FDR. Although identifying similar numbers of loci as the
518 generation 22 analyses, P-values for individual SNPs for generations 22 and 61 show
519 little similarity (Figs 2B-C), with arcsine-square transform Pearson's r = 0.116.
520 Ultimately, the SNPs identified at generations 22 and 61 were largely different.
521 Moreover, the SNPs identified at generation 61 clustered into only 11 unique regions, as
522 compared with the 258 regions for generation 22 (Fig 3A-B).
523
524 **Fig 2. P-value comparisons between generations.** Scatterplot comparisons of the
525 generations 22 and 61 p-values with Pearson's correlation: (A) Generation 22
526 regularized T-test vs generation 22 WRT test (cor = 0.9997). (B) Generation 22 WRT
527 test vs generation 61 WRT test (cor = 0.0909). (C) Generation 22 WRT test vs
528 generation 61 WRT test (cor = 0.1156). (D) Distribution of raw p-values (generation 22).
529
530 **Fig 3. Manhattan plots and volcano plots of differentiation analyses.** Manhattan
531 plots for results from (A) the generation 22 WRT test (shared loci), (B) the generation 61
532 (pooled) WRT test (shared loci). The red peaks indicates those that exceeded critical
533 threshold (FDR = 0.01) for that individual test. Volcano plot including -logP vs HR allele
534 frequency minus C allele frequency for (C) generation 22 and (D) Generation 61
535 (orange points indicate HR AF > C AF; green points indicate HR AF < C AF). (E)
536 Scatterplot comparing the -log p-values of the generation 61 mixed model analyses
537 (individual mouse) with -log p-values produced when these same data are treated as
538 pooled sequencing allele frequencies and analyzed with WRT test. Red line has
539 intercept = 0 and slope = 1. Green line represents the least squares regression line.

Table 1. WRT results.

Data	Total Loci	FDR 0.01 (-logP)	Significant SNPs	All Regions	Regions after strict culling
Gen22AF	2,045,546	2.66	630	187	6
Gen61AF	2,045,546	3.06	1,285	11	4
Gen22AF	4,446,523	2.62	1,184	258	13
Gen61AF	5,932,148	3.23	1,449	11	5

Number of SNPs listed represents those that are statistically significant based on a False Discovery Rate of 1% using permutations. Analyses with 2,045,546 loci incorporate only loci which are shared between generations 22 and 61. Regions distinguished by being separated from the next closest significant locus by more than 1 million bp. Additional regions remaining after additional culling methods have either 20 significant loci or at least 2 significant loci with one having a p-value <1.00E-04.

Given such notable differences between the SNPs and regions implicated by generation 22 and 61 analyses (Table 2), analyses were repeated focusing only on the loci found in both data sets (N = 2,045,546). With fewer loci being analyzed, fewer significant SNPs were identified at FDR = 0.01, as well as fewer regions for all analyses except for WRT with generation 61. The total peaks identified when using only the shared SNPs includes 187 and 11 regions for generations 22 and 61, respectively.

Table 2. Genomic regions identified as differentiated under "strict" culling

G22 Region	G22 WRT	G61 Region	G61 WRT	Chr	minPOS	maxPOS	Size	Loci	Shared Loci
1	x			1	152,318,219	153,239,876	921,658	40	25
2	x			2	78,021,909	78,974,325	952,417	3	0
		1	x	3	51,199,110	51,602,693	403,584	124	65
3	x			5	32,384,612	32,975,871	591,260	32	4
4	x			5	102,846,390	106,315,986	3,469,597	63	37
		2	**x**	**6**	**40,933,658**	**41,748,676**	**815,019**	**5**	**1**
5	x			6	122,815,876	124,446,843	1,630,968	43	20
		3	**x**	**9**	**41,413,436**	**42,478,817**	**1,065,382**	**1,277**	**647**
6	x			9	80,349,989	82,894,555	2,544,567	27	20
7	x			10	14,067,617	18,376,599	4,308,983	25	7
8	x			10	20,890,526	21,419,406	528,881	33	4
		4	x	10	104,966,751	105,529,701	562,951	2	0
9	x			13	46,088,694	46,866,721	778,028	33	13
10	**x**			**14**	**52,115,206**	**53,776,455**	**1,661,250**	**42**	**7**
11	x			14	77,333,032	78,080,942	747,911	4	3
		5	**x**	**15**	**19,245,017**	**20,197,326**	**952,310**	**27**	**17**
12	x			18	57,707,454	60,118,834	2,411,381	128	81
13	x			18	78,018,740	78,504,680	485,941	4	4

A test is deemed to have produced a differentiated region if the region contains at least 20 SNPs significant at FDR = 0.01 or at least 2 SNPs significant at FDR = 0.01 and at least one SNP with p-value < 1.0E-04 (See Methods: Determination of Selection Signatures). "Loci" listed are those significant at FDR = 0.01 (Table 1), the counts themselves match the number of differentiated loci identified by the statistical test with the most such loci. "Shared Loci" are the number of loci listed in the "Loci" column that are also shared between both generations.

Bolded loci match "consistent" regions identified by Hillis et al. (2020)

Regions after "strict" culling

Using all available SNPs for generation 22, after applying "strict" culling (see Methods), the remaining regions were reduced to 13. All of the regions implicated by these analyses included or were near genes with intuitive implications for running behavior (see Discussion). For generation 61, strict culling reduced the total peaks to only 5 unique regions.

Despite the HR lines reaching selection limits around generation 22 or shortly thereafter (Careau et al. 2013), the most differentiated 13 regions (Table 2) have little fixation. Of the SNPs in these regions (N =79,198), only about 8.78% are fixed in the HR lines, which is not significantly different from the 9.21% fixed in the control lines (unequal variance t-test comparing % fixed in the 4 HR versus 4 C lines: p-value = 0.4322). If we repeat this fixation comparison for the loci shared between generations 22 and 61 (N = 42,745), 1.62% are fixed in the HR lines, which is still not significantly different from the 1.58% fixed in the control lines (p-value = 0.6129).

Comparison of selection signatures at generation 61 for individual vs. pooled sequencing data

Originally, the generation 61 individual mouse data were analyzed using mixed models (Hillis et al. 2020). We compared the previously published p-values from those analyses with the p-values produced after pooling data by line and analyzing by the WRT test (Fig 3e). The mixed model analyses produced lower p-values in general, as would be expected due to loss of power with pooling (Xu and Garland 2017), with the difference being greater for lower p-values. As a result, fewer SNP loci and hence

588 fewer chromosomal regions were identified as significantly differentiated between the
589 HR and C lines with pooled data. Of the total regions detected with FDR = 0.01, 7 were
590 identified at generation 22 that matched the 13 “consistent” regions identified with the
591 mixed model analyses (Hillis et al. 2020). The 6 consistent regions that were not
592 identified by analyses of the pooled data tended to have relatively large p-values for
593 individual SNP loci or cover a narrower area of the genome, as compared with the other
594 7 consistent regions.
595

596 ## Divergence over time

597 The 5th percentile threshold for the difference in T-scores determined by permutations
598 was 5.139787. About 6.44% of the T-scores for unpermuted data were larger than this.
599 This difference of 6.44 - 5% indicates that ~1.44% of our values for the real data
600 (approximately 28,470 SNPs) may be considered nominally statistically significant for α
601 = 0.05. This result provides statistical support for our claim that the selection signatures
602 differ between generations 22 and 61 (see Discussion). Defining a region as containing
603 at least 20 significant SNPs with no adjacent SNPs separated by more than 1mbp, and
604 considering the 1,400 most significant SNPs, they cluster into 14 regions on 13
605 chromosomes (Table 3).
606

Table 3. Genomic regions of divergent evolution

Chr	minPOS	maxPOS	Loci	Median_T	Highest_T	T_position	g22	g61[a]
1	152,255,010	153,208,066	44	48.9	94.9	152,795,939	WRT	
1	156,267,494	156,908,946	21	48.5	115.5	156,699,891		
2	153,709,397	157,032,971	20	46.0	102.8	154,908,418		1 Test
3	45,831,668	52,497,670	49	51.0	139.9	51,543,977		2 Tests
4	89,020,582	90,615,570	26	52.5	103.3	90,007,699		1 Test
5	107,675,741	111,271,760	30	55.2	232.2	109,810,077		Consistent
9	41,416,364	42,248,169	31	47.3	99.8	41,533,058		Consistent
10	101,702,890	105,671,151	154	47.4	177.0	103,108,543		2 Tests
12	109,050,203	110,779,901	28	46.6	109.2	109,228,613		
14	96,560,689	98,613,561	76	49.5	212.2	97,831,005		Consistent
15	18,635,736	20,608,793	39	51.4	208.5	19,984,048		Consistent
16	45,132,582	47,948,007	23	46.5	163.9	45,158,346		
18	69,603,969	74,277,731	62	46.7	147.5	73,016,958		
19	35,121,427	35,736,790	40	48.7	250.1	35,699,388		

[a]Tests here are the three tests used by Hillis et al. 2020 (local maxima, haplotype, and FixedHR/PolyC). "Consistent" is the term used to describe regions identified by all three tests.

Simulations to compare presumptive statistical power across generations

Simulations were conducted to gauge how much the allele frequencies determined through sequencing reflect allele frequencies of the actual populations at generations 22 and 61. Generation 22 allele frequencies have greater variance from the actual population AF than generation 61 (see Figs 4A-B for an example of the 0.5 population AF distribution). The greater error variance in generation 22 is associated with reduced statistical power of 0.3864 versus 0.5031 for generation 61 when comparing simulated allele frequencies of 0.4 and 0.6 (Figs 4C-D).

Fig 4. Variance and power simulation results. Simulations for a population allele frequency of 0.5 for (A) generation 22 and (B) generation 61. See Methods for details. Values shown are allele frequencies for each of 100,000 simulated data sets for a single line. Methodological differences in the sampling of mice and sequencing procedures for the two generations result in greater sampling error for generation 22 (i.e., larger SD). Note that binning is done such that loci that fall on a break (e.g., 0.05) are grouped into the lower bin (e.g., 0 to 0.05). Similar simulations were then conducted to create data sets for use in estimating statistical power for detecting selection signatures for generations 22 and 61. (C) Distribution of p-values for simulated allele frequencies of 0.4 versus 0.6, for generation 22. Power for $\alpha = 0.05$ is 0.3864. (D) Distribution of p-values for simulated allele frequencies of 0.4 versus 0.6, for generation 61. Power for $\alpha = 0.05$ is 0.5031.

Simulations comparing power with and without a biological constraint

Simulations were performed modeling response to selection assuming either a constraint with a base of 10,000 revolutions per day and the capacity to evolve to about 17,000 (in the winter) or no such constraint (see Methods). For both constrained and unconstrained simulations, wheel running for HR and control lines diverge recognizably at least by generation 6 (Fig 5A), consistent with the selection experiment. The replicate HR lines for unconstrained and constrained models appear fairly similar for earlier generations (Figs 5B-C), presumably because mice are not widely achieving constrained running levels. As expected, the among-line variation for control lines increases gradually across generations. For the HR lines, among-line variance does not increase to a noticeable extent and potentially even diminishes by later generations, a result that is also consistent with the selection experiment (Garland, Jr., Kelly, et al. 2011).

Fig 5. Power simulations considering a constraint. Simulated running levels for (A) mean running of 4 non-selected control lines (blue) for 200 simulations compared with mean running of 4 HR lines under 100 unconstrained simulations (dark red) and 100 constrained simulations (light red). (B) Individual HR and control lines single unconstrained simulation. (C) Individual HR and control lines single constrained simulation. Black lines show the mean narrow-sense heritability for four lines within each linetype for (D) 100 control line simulations (arbitrarily the first 50 from each model), (E) 100 simulations for HR lines with unconstrained running, and (F) 100 simulations for HR lines with constrained running (16,000 revolutions). Standardized

657 selection differentials (calculated within family and sex) from simulations for (G) control
658 lines, (H) HR lines without a constraint, and (I) HR lines with a constraint. The red line
659 represents the mean of all heritability and selection differentials (N = 100). Note
660 different axes for panel G versus H and I.
661
662 The calculated heritability (slope of the regression of offspring [generation 1] on
663 midparent [generation 0]) for all 200 simulations for control lines indicate that our
664 parameters resulted in a narrow-sense heritability of about 0.3621 (N=8,000 families).
665 For individual lines, the estimated heritability for successive generations was highly
666 variable, as would be expected with such small sample sizes (10 families/line).
667 However, the means clearly indicate a slow loss of heritability in the control lines and a
668 more rapid loss in the HR lines, although values never go to zero (Figs 5E-F),
669 consistent with the selection experiment (Careau et al. 2013).
670 The standardized selection differentials (calculated within family and sex) for the
671 unconstrained model remained very consistently around 0 for the control lines (Fig 5G)
672 and 1.2-1.3 for the HR lines (Figs 5H-I). However, the constrained selection differential
673 is on average 0.016 below the unconstrained differential. Although slight, this difference
674 remains consistent across nearly all generations (graph not shown). In the actual
675 selection experiment, selection differentials declined across generations (Careau et al.
676 2013).
677 Under both models, Type I error rate for $\alpha = 0.05$ when comparing allele
678 frequencies of HR with C lines was deflated at generation 0, regardless of the effect
679 size for the locus. Type I error ranged from 0.0313 to 0.0420 with no preference for any

680 effect size (S1 Table). This relatively low power when comparing the HR and control
681 lines (when the line itself is the experimental unit) has been documented previously with
682 simulations for both genetic data and phenotypes (Hillis et al. 2020; Castro et al. 2021).
683 As expected, power to detect differentiation between the HR and C lines
684 increased across generations, but never exceeded 0.057 for any generation for either
685 model. Comparing models at each generation indicates that power is significantly
686 higher under the unconstrained model by generation 15, although the difference is trivial
687 (0.0015 with P=0.0127) (Table 4). This differential in power increased through
688 generation 50, when it reached 0.0055 (P=2.63E-15), before beginning to diminish with
689 later generations. Although the information in Table 4 does not tell us about the power
690 to detect loci based on effect size (see numbered list below), it does establish that we
691 expect more total selection signatures at generation 61 than 22 (see Discussion).
692
693 **Table 4. Simulated statistical power for constrained and unconstrained models**

Generation	Unconstrained power	P-value for Unc. Comparing present generation to previous generation[a]	Constrained power	P-value for C. Comparing present generation to previous generation[a]	P-value for Constrained vs Unconstrained[b]
0	0.0408[c]	NA	0.0409[c]	NA	0.8436
5	0.0417	0.1192	0.0423	0.0365	0.3688
10	0.0429	0.0379	0.0429	0.2685	0.9817
15	0.0447	0.0060	0.0432	0.6693	0.0127
20	0.0456	0.1272	0.0425	0.1891	2.69E-07
22	0.0461	0.4333	0.0425	0.9470	2.31E-07
25	0.0465	0.5290	0.0434	0.1721	1.58E-06
30	0.0484	0.0042	0.0447	0.0374	4.61E-08
35	0.0503	0.0041	0.0463	0.0155	4.51E-09
40	0.0527	0.0006	0.0476	0.0423	5.31E-13
45	0.0544	0.0160	0.0497	0.0009	6.42E-11
50	0.0557	0.0584	0.0502	0.4380	2.63E-15
55	0.0562	0.4050	0.0510	0.2169	2.40E-13
60	0.0565	0.7307	0.0512	0.8162	3.70E-13
61	0.0563	0.8387	0.0512	0.9147	6.27E-13

694
695 [a] These p-values are calculated using a T-test assuming unequal variance comparing the power
696 of the generation for that line to the previously listed generation (e.g., unconstrained power at
697 generation 5 compared to generation 0 has a p-value of 0.1317).
698 [b] P-value from a T-test (unequal variance) comparing the power of the 100 unconstrained
699 simulations to the 100 constrained simulations
700 [c] Type I error rate.

701

702 The average Pearson correlation between p-values across 2,096 loci for
703 generation 22 and 61 for the constrained model ($r = 0.3843$) was not statistically
704 different from that for the unconstrained model ($r = 0.3889$: total N = 200, unpaired-T = -
705 1.6938, P = 0.0919). In the unconstrained model, 33.3% of the loci significantly
706 differentiated at generation 22 ($\alpha = 0.05$) were still differentiated at generation 61,
707 versus 31.5% under the constrained model (unpaired-T of percentage of loci
708 consistently different for generations 22 and 61 at $p \leq 0.05$ for 200 simulations = -
709 2.4269, P = 0.01614). This consistency of about 1/3 is more than 3 times greater than
710 for the real data (9.12% for T-tests), which mirrors the drop in the correlation of p-values
711 between generations 22 and 61 (Fig 2B, $r = 0.0898$). Incorporation of sampling error
712 into the constrained model lowered the correlation between p-values to 0.2695 and the
713 proportion of loci significant at generation 22 still significant at generation 61 to 22.7%.
714 Comparisons of power to detect differentiation between the HR and C lines in
715 relation to effect size of locus and generation under two simulation models (S2 Table)
716 indicates:
717 1. power increased with effect size, as expected;
718 2. at generation 22, power was greater in the unconstrained model for loci with effect
719 sizes 25.6 or above;

3. at generation 61, power was greater under the unconstrained model for loci with effect sizes 12.8 and above (excluding the largest effect size, 204.8);
4. under both models, power was consistently greater at generation 61, except for the largest effect size, where the power is reversed.

Comparison of selection signatures in generations 22 and 61 for pooled data

When the average allele frequencies of SNPs within regions identified by strict culling at generation 22 (this study) are compared to the average AF of those loci at generation 61, an increase in among-line variance is apparent for generation 61, within both the HR and C linetypes (Fig 6A). All else being equal, this increase in among-replicate variance should lower the statistical power to detect differentiation between the HR and C linetypes. In agreement with this expectation, most of these strict regions at generation 22 (Table 2) are no longer significantly differentiated at generation 61 (Table 2 and Fig 6A). However, several of the 5 strictly culled regions at generation 61 also show some evidence of differentiation at generation 22 (Fig 6E). Although strict culling methods exclude regions identified with generation 61 AF analyses, regions implicated in generation 22 with FDR = 0.01 culling alone do have considerable overlap with some of these 5 regions identified at generation 61. Generation 61 regions 3, 4, and 5 (Table 2) were significant at FDR = 0.01 for all three analyses at generation 22 (S3 Table).

Fig 6. Power simulations considering a constraint. (A) Allele frequencies of regions identified as significant (via strict culling) at generation 22 (Table 2) (excepting region 18, for which no loci were available in the generation 61 data). Line plot examples are

743 provided for generation 22 (B) region 1, (C) region3, and (D) region 10, with generation
744 22 on left and generation 61 on right. (E) Allele frequencies of regions identified as
745 significant at generation 61. (F) Line plot example for generation 61 region 6. SNPs
746 included have a nominal $p<0.05$ at generation 22 and any SNPs at generation 61 which
747 matched the generation 22 SNPs (shared loci, see Table 1).
748
749 For some of the regions identified as significant at generation 22, differentiation
750 may have been lost by generation 61 as result of a single line diverging from the others
751 (for example, line 3 in region 1 or line 7 in region 3 [Figs 6B-C]). In general, mean
752 differences at generation 22 are much smaller than at 61, but also with much less
753 among-line variance. A particular example of this includes region 10 (Fig 6D), which is
754 the only region identified at generation 22 (after strict culling) to continue to be detected
755 as differentiated at generation 61 (see Discussion).
756

757 Possible biological function of generation 22 differentiated regions

758 A total of 79 genes (including predicted genes and miRNA) were identified using "strict"
759 culling of generation 22 regions. These were insufficient for powerful ontology tests and
760 so regions containing at least five differentiated loci at FDR <0.01 were included,
761 bringing the total number of included genes to 345 (S4 Table). Of these 345
762 differentiated genes, 285 were recognized by the Panther database (used by the Gene
763 Ontology Resource) and used for identifying potential biological function. Those not
764 recognized were generally miRNA, predicted, or olfactory genes.

GO Biological Functions implicate antifungal innate immune response, sensory perception of smell, and embryonic skeletal system morphogenesis. The antifungal enrichment appears to be the result of a group of C-type lectin genes found on chromosome 6 (chr6:122,815,876-124,446,843). It may not be a coincidence that this region also includes a group of vomeronasal genes contributing to the sensory perception of smell term. The genes that implicate the embryonic skeletal morphogenesis term include a cluster of *Hoxb* genes found on chromosome 11 (chr11:93,129,916-96,570,699).

A few genes specifically from the 79 found in the "strict" regions merit mention for their relevance to the running phenotype, including *Cited2* (adrenal cortex formation), *Rbm24* (positive regulation of skeletal muscle fiber differentiation), and *Dspp* (negative regulation of bone development).

Discussion

Overview

Previously, whole-genome sequence data for individual mice at generation 61 of the High-Runner mouse selection experiment were used to identify genomic regions differentiated between HR and control lines. Thirteen of these were termed "consistent" because they appeared with three different analytical methods (Hillis et al. 2020). These 13 regions contained genes associated with known phenotypic differences between the HR and control lines and intuitive associations with running ability and/or motivation/reward systems. However, given that the HR lines had begun to reach selection limits around generations 17-27, depending on line and sex (Careau et al. 2013), tens of additional generations, with continuing random genetic drift, could have obscured many selection signatures. Therefore, in the present study, we analyzed allele frequencies for the lines sampled at generation 22, based on DNA pooled by line. These analyses of generation 22 identify hundreds of genomic regions differentiated between the HR and C lines (FDR = 0.01), despite using pooled sequence data rather than sequences for individual mice (Xu and Garland 2017). We then reanalyzed the data from generation 61 as allele frequencies by line, to mimic the data available for generation 22, and found that the regions identified as differentiated at generation 61 are, at best, weakly differentiated at generation 22. Nevertheless, both generations' differentiated regions contain genes that make biological sense for wheel-running behavior. Below, we discuss (1) implications of the differences in data type between generations 22 and 61, (2) possible statistical and biological explanations for the

804 differences in identified regions, and (3) genes and biological systems highlighted by the
805 genomic regions identified by generation 22 analyses (after strict culling).
806
807 ## Differences in selection signatures at generations 22 and 61
808 We expected estimates of selection signatures to be similar at generations 22 and 61,
809 based on the fact that the HR lines had mostly reached selection limits by generation 22
810 (Careau et al. 2013), such that the most biologically important loci would have gone to
811 fixation or at least reached equilibria across most or all of the HR lines. In agreement
812 with this expectation, of the 13 "consistent" regions identified by Hillis et al. (2020) for
813 generation 61 (using individual mouse data), 8 were still identified by at least one of the
814 tests (FDR = 0.01) using the generation 61 genotypes pooled into allele frequencies per
815 line. Generation 22 analyses of pooled sequence data identified 7 of the 13 consistent
816 regions (although several of these regions were only detected by a few SNPs: S3
817 Table). Interestingly, the consistent region on chromosome 14 was more strongly
818 detected at generation 22 than at generation 61 using pooled sequence analyses (Table
819 2).
820 On the other hand, the strongest selection signatures observed at generation 61
821 with the data treated as pooled sequences are not among the strongest ones observed
822 at generation 22 (based on number of SNPs detected and their p-values), despite
823 continued selection on the HR lines. For example, region 1 of the generation 61 “strict”
824 culling pooled analyses (chr3:51,199,110-51,602,693) included 124 SNPs (Table 2),
825 and whereas generation 22 analyses did not detect any of these loci as significant.
826 Another example, region 3 of the generation 61 pooled analyses (chr9:41,413,436-

42,478,817) included 1,277 SNPs (FDR = 0.01), but none of the 647 shared loci were identified by in the generation 22 analyses (S3 Table). When directly comparing SNPs differentiated at FDR = 0.01, we see only a single SNP of overlap for the WRT test.

In addition to the differences in individual SNP results, a 17-fold greater number of regions was identified by generation 22 analyses than generation 61 pooled analyses at FDR = 0.01 (Table 1). This ratio applies to all statistical tests and the complete SNP analyses for each generation, as well as the analyses of SNPs shared by the two generations. Moreover, the SNPs identified at generation 61 were clustered into far fewer regions (Table 1). Broadly, this difference in numbers of selection signatures have at least two possible explanations, which are not mutually exclusive: (1) differences in data type, quality, and quantity; (2) biological differences between generations 22 and 61.

1) Differences in data type, quality, quantity, and sampling error

Our power to detect differentiation in allele frequencies should have been lower for generation 22 than for generation 61 (Figs 4C and D). As also noted in the Methods, the estimates for SNP allele frequencies per line at generation 61 were based on ~10 mice/line sampled and an 12X average read depth per mouse, yielding a total of 5,932,148 variable SNP loci (Hillis et al. 2020). For generation 22, pooled sequencing was done with ~20 mice/line and an average read depth of 24X, yielding 4,446,523 variable SNPs (Table 1). Generally, with an average read depth of 12X per mouse, both alleles will be represented for each mouse (i.e., 20 alleles per line) for generation 61 allele frequencies. However, with 24X average read depth for generation 22,

simulations involving sampling alleles with replacement show that generation 22 is prone to vary more from the actual population allele frequency (Figs 4A and B). Thus, the much greater number of differentiated SNPs and chromosomal regions detected at generation 22 would not appear to be simply a function of greater statistical power versus generation 61. Thus, we now consider possible biological explanations.

2) Biological differences

One way to highlight the differences in selection signatures detected at generations 22 and 61 is to note that of the differentiated regions detected for generation 61, two of them contain hundreds of statistically significant SNPs (FDR = 0.01) shared between the generation 22 and 61 data sets. Despite this, those two regions are not among the more differentiated regions in the Manhattan plots (Fig 3A and B).

What biological explanations might account for such discrepancies? One possibility is a physiological constraint that eliminates the need for all loci favorable to wheel running to be maintained at high frequencies once a selection limit is reached (see verbal model in Hillis and Garland, Jr. 2022). We consider this from the perspective that many complex traits are influenced by hundreds or thousands of loci (Wood et al. 2014; Long et al. 2015). Voluntary exercise behaviors would likely be among them, given that they incorporate numerous physiological and morphological traits related to ability (e.g., cardiac muscle, skeletal muscle, bone, metabolism, water and temperature homeostasis) as well as aspects of motivation and reward (e.g., dopamine signaling, chemosensory systems) (Lightfoot et al. 2018; Wang et al. 2022).

872 Although biological constraints can be defined in various ways (Garland et al.
873 2022), in the present context, a constraint would be anything that limits the maximum
874 revolutions that an individual mouse can run during the testing period. Previously, we
875 discussed how different unique responses to identical selection criteria (i.e., “multiple
876 solutions”) could occur and referenced constraints as a potential explanation (Hillis and
877 Garland, Jr. 2022). To utilize and expand on their example, suppose that mice are
878 subject to a constraint on wheel running caused by joint pain: they stop running when
879 the pain becomes intolerable. In this scenario, joint pain is sufficient to limit wheel
880 running and it serves as a “weak link” or single limiting factor in the biological systems
881 required for high wheel running. Then suppose 10 alleles located at 10 independent
882 biallelic loci, with entirely additive effects, are capable of increasing wheel running.
883 Suppose further that only five such alleles are needed to achieve the maximum amount
884 of wheel running permitted by joint pain. Under this model, if selection acts on a
885 population to increase running, then (1) fixation of the favorable allele at any five of the
886 10 loci will coincide with a selection limit determined by pain tolerance, (2) none of the
887 alleles at any of the 10 loci must be fixed to reach the pain-determined limit, (3) more
888 than 5 favorable alleles could be maintained at intermediate allele frequencies, and (4)
889 as long as enough favorable alleles are maintained for the selection limit, some
890 favorable alleles can be lost without detriment to wheel running. These factors allow for
891 substantial variation among the replicate lines and considerable flexibility for change
892 within a given line, even for favorable wheel-running alleles at the selection limit. This
893 possibility of "genetic churn" beyond a selection limit that is caused by a physiological
894 constraint also implies that genotype-to-phenotype maps (Travisano and Shaw 2013;

Zamer and Scheiner 2014; Porto et al. 2016; Zinski et al. 2021) may be moving targets and hence difficult to identify. Therefore, we used simulations to compare power to detect and consistency in detected selection signatures, both with and without a physiological constraint.

Allele frequency divergence over time

The relatively large number of SNPs identified to be divergent between generations 22 and 61 illustrate a shift between the generations. Furthermore, regions with the greatest divergence between the generations align closely with some of the previously identified regions, particularly those identified by the generation 61 mixed model analyses (Table 3). By generation 22, the lines had not had as much time to evolve as much separation in allele frequencies between HR and C as by generation 61. As a result, the differences between the linetypes are generally between -0.5 and 0.5 at generation 22 (Fig 3C), whereas this difference expands to between -1 and 1 for generation 61 (Fig 3D). Given that the significant regions identified at gen 61 are typically going to be those whose difference is near -1 or 1, most likely those same regions at gen 22 had differences within the -0.5 and 0.5 range. Therefore, a growing difference in HR and C allele frequency had to have occurred, in our data, between gen 22 and gen 61 to observe a significant difference at 61. Such loci would naturally be among the more significant in the difference of t-tests over time.

Simulations comparing power under constrained versus unconstrained models

These simulations were conducted to test whether a physiological constraint on the phenotype of wheel-running behavior could reduce the consistency of loci identified at different generations of a selection experiment. To better simulate realistic phenotypic variance within the population, both the wheel-running and constraint phenotype simulations were based on equations with both genetic and environmental sources of variance, such that both could evolve.

Similarities between the constrained model and real data

The constrained model appears to better reflect what we observe in the actual response to selection. This is due to the lack of a clear selection limit achieved under the unconstrained model (Fig 5A). Although the response to selection diminishes over time (likely due to the reduction in heritability: Fig 5E), a clear plateau is not apparent. In contrast -- as must be the case -- a clear plateau occurs under the constrained model.

Correlation between generations 22 and 61 p-values

For the tests comparing allele frequencies at each of 2,096 loci between the HR and C lines, the correlation of p-values between generations 22 and 61 was statistically lower in the constrained model (r=0.3843) as compared to the unconstrained model (r=0.3889), though still 4x higher than for the real genomic data (r=0.0909) (Fig 2). Even with the inclusion of sampling error, the correlation (r=0.2695) is nearly 3x greater than for the real genomic data, which indicates that other factors (e.g., gene interactions) must be contributing to the differences between the generations.

939 In spite of the similarity in the correlation of p-values between generations for
940 simulations, the between-generation consistency of detected selection signatures was
941 slightly but statistically greater under the unconstrained (33.3%) than under the
942 constrained model (31.5%). This difference may be due to the constrained model
943 having very slightly (~0.016) though consistently lower selection differentials (Figs 5G
944 versus 5H), which could lead to less fixation of favored alleles. However, the relatively
945 small difference between models in consistency of selection signatures is not enough to
946 explain the large differences in the real data between generations 22 and 61 (Table 2).
947 The inclusion of sampling error into the estimates decreased the 31.5% consistency
948 between generations 22 and 61 differentiated loci to 22.7%. This level of consistency
949 with simulated data remains more than 2-fold higher than for the real data (9.12%), thus
950 implicating the presence of additional factors that reduce consistency in the real data
951 (e.g., epistatic effects).
952 Overall, our simulations fail to demonstrate why we observe a 17X drop in
953 significant regions from generation 22 to 61 (Table 1), implying instead that we should
954 detect more at generation 61 than at 22 (Table 4).
955
956 ## Effect sizes of loci
957 Under both models, more loci were detected at later generations (Table 4). However,
958 the power to detect loci with the largest effect size was much higher at earlier than later
959 generations (S2 Table and S1 Fig). This pattern makes sense in consideration of the
960 factors that affect the average difference in allele frequency between the HR and control
961 lines and the variance among replicate lines within linetypes. Drift will generally

962 increase the variance among lines with each generation. The allele frequencies in the
963 simulated control lines will be affected only by this drift. Allele frequencies in the HR
964 lines will be affected both by drift and selection, where selection will have stronger
965 effects at loci with larger effect sizes. This results in something of a race between
966 selection increasing the difference in allele frequencies between HR and control lines,
967 while drift increases among-line variance for both HR and control lines. For loci with
968 small effect sizes, drift will have a relatively greater influence over allele frequencies
969 than selection at any generation, and thus detection rates never vary far from the Type I
970 error rate, i.e., power is virtually zero (S2 Table). Loci with large effect sizes, however,
971 are able to differentiate rapidly, often leading to fixation of the favored allele in our
972 simulations (S1 Fig). Even after fixation in the HR lines, drift is still able to increase
973 allele-frequency variance among the control lines (potentially to the point of fixing loci
974 for opposite alleles), thus further reducing the power to detect any differentiation. Thus,
975 the power to detect signatures of selection should increase the most rapidly across
976 generations for loci with the largest effect sizes, but power is also expected to decline
977 after fixation of the favored alleles in the HR lines and with continuing increase in
978 variance among the control lines (S1 Fig).
979 That the power to detect a locus as differentiated is correlated with its effect size
980 is unsurprising. For example, under the unconstrained model the power to detect
981 selection signatures for loci with 0.4 effect size is about 16.6-fold less than the power to
982 detect loci with 204.8 effect size and 12.8-fold less for the constrained model
983 (generation 22). This gap diminishes to about 7.7-fold difference (both models) by
984 generation 61 (S2 Table), presumably due to the reasons described in the previous

985 paragraph. However, the 0.4 effect size loci are far more numerous than the 204.8
986 effect size loci (N = 720 and 8, respectively). Consequently, the number of 0.4 effect
987 size loci detected as significant is nearly 5-fold greater (unconstrained) and more than
988 7-fold greater (constrained) than the number of 204.8 effect size loci detected. The
989 most notable difference between the constrained and the unconstrained models is that
990 at generation 22 the unconstrained model yielded substantially more power than the
991 constrained model for loci with the largest effect sizes (0.724 to 0.485, respectively;
992 unpaired t-test, P=4.01E-22). This would imply that constraints may have a substantial
993 impact on the ability to detect selection at loci with the greatest effect sizes, a result that
994 deserves further study.
995 For identifying possible biological functions, we would ideally focus on loci with
996 relatively large effect size, as these will have the most direct influence on the phenotype
997 and may serve as potential targets for future functional studies. We have no information
998 on effect sizes of SNPs or regions detected as differentiated for our real data. The
999 relative proportions of low- and high-effect size loci among the detected selection
1000 signatures in the real data will likely vary from our simulations, depending on the actual
1001 distribution of those effect sizes and other factors. However, the simulations do suggest
1002 that we may have numerous small-effect size loci among our detected selection
1003 signatures. The inclusion of the “strict” culling method was meant to prioritize regions
1004 that would have large effect sizes. Having more loci that are differentiated and linked
1005 together would be expected from those regions under strong selection because
1006 recombination would have fewer generations to break up linked base pairs before the
1007 region becoming fixed in the HR lines. As the simulations have so many more loci with

small effect sizes, at generation 0, when we compare the lowest p-value produced for each simulation for the 0.4 effect size we tend to see lower p-values than loci with 204.8 effect size simply because of more opportunities to produce a low p-value. However, generation 22 appears to be better for detecting a greater proportion of selection signatures from loci with large effect sizes as the relative proportion on large effect size loci appears to be higher (S2 Table).

Possible biological functions of generation 22 differentiated regions

Ontology analyses identified biological processes that can be grouped into three categories: sensory perception of smell, antifungal innate immune response, and embryonic skeletal system morphogenesis. Of these, the system that is most consistent between generations 22 and 61 is the perception of smell, which was among the mostly clearly differentiated systems at generation 61 (Hillis et al. 2020). As was discussed by Hillis et al. (2020), the experimental procedure for measuring wheel running, for logistical reasons, involved mice being placed on wheel over three batches and mice in batches 2 and 3 are placed on wheels which still smell of the previous mouse (Swallow, Garland, Jr., et al. 1998). Evidence of an evolutionary response to this is visible in the HR lines in that HR mice will run at very different speeds if on a wheel that is clean, previously traversed by a male, or previously traversed by a female (Dewan et al. 2019). Alterations in the transcriptome also indicates changes in olfactory and vomeronasal systems (Nguyen et al. 2020). Taken together, these results indicate that perception of smell may be a notable factor in their motivation for running on the wheels and also consistent with the idea that motivation is expected to evolve before

ability (Garland, Jr. et al. 2016; Khan et al. 2024). Interestingly, although both generations demonstrate evolution in genomic regions association with olfaction and vomeronasal, the regions implicated in each generation are different, with an exception of the region on chromosome 14 (chr14:52,115,206-53,776,455), which was identified by the generation 22 WRT analyses and the generation 61 mixed model analyses (Table 2). However, additional studies should be done to address the effects of olfactory/vomeronasal systems more directly on running behavior of the HR mice. This could be done with ablation procedures on HR and C mice and observing changes in running behavior. The antifungal ontology term is possibly a hitchhiker with the vomeronasal genes also present in the differentiated region (chr6:122,815,876-124,446,843).

Ontology analyses also indicated embryonic skeletal system development as result of a group of *Hoxb* genes within a differentiated region. If these *Hoxb* are the driving force underlying the many skeletal differences that have been documented between HR and C lines (Garland, Jr. and Freeman 2005; Kelly et al. 2006; Middleton et al. 2008; Middleton et al. 2010; Wallace et al. 2010; Wallace et al. 2012; Castro and Garland, Jr. 2018; Copes et al. 2018; Schwartz et al. 2018), then this is an exciting discovery because it would represent a response to selection in a group of genes known to be evolutionarily influential in body patterning and development (Stratford et al. 1999). However, whether the *Hoxb* genes are the cause of skeletal differentiation is unclear. Although *Hoxb* genes may play a role in these changes, they are far from the only candidates. GO term “skeletal system development” includes 7 additional non-*Hox* genes, including *Phospho1*, *Col1a1*, and *Mbtd1*, which are all located in the same

differentiated region as the *Hox* genes. Furthermore, individual loci demonstrating the greatest differentiation do not appear to be in *Hox* genes themselves or their regulatory regions. Even if *Hox* genes are a hitchhiker in a region with other genes more directly targeted by selection due to their skeletal effects, exploring potential side effects of this evolution would be of interest. Expression analyses during developmental stages when these genes are most active may provide insight into how *Hox* genes may be altered in the HR mice.

Other genes of potential interest

The 79 genes included in top regions also contain a few of particular note: *Cited2*, *Rbm24*, and *Dspp*. Each of these genes is associated with ontologies and phenotypes that have been identified as differentiated between the HR and C mice. *Cited2* is a gene whose knockout (KO) has been associated with alterations in brain and heart morphology (Barbera et al. 2002; Bamforth et al. 2004; MacDonald et al. 2008) and has also been associated with adrenal development (Val et al. 2007). As noted in the introduction, HR mice have larger brains and hearts than C mice (Kolb et al. 2010; Kolb, Rezende, et al. 2013; Kolb, Kelly, et al. 2013; Kelly et al. 2017). Additionally, adrenal corticosterone levels were found to be different between the linetypes (Malisch et al. 2007; Garland, Jr. et al. 2016). *Rbm24* is a gene associated with skeletal muscle fiber differentiation, particularly during regeneration following injury (Cardinali et al. 2016; Zhang et al. 2020; Grifone et al. 2021). The HR and C lines have demonstrated differences in muscle fiber types within muscles important for wheel running such as the gastrocnemius (Syme et al. 2005; Guderley et al. 2008; Castro et al. 2022). However,

differential response to muscle injury has not been found between the linetypes (Kay et al. 2022). Lastly, *Dspp* was identified among the differentiated genes. This gene has been associated with development of long bones (such as femurs) and cortical and trabecular bone thickness (Verdelis et al. 2008; Jani et al. 2016). The HR and C mice have shown various differences in bone morphology (see Introduction).

Limitations and conclusions

Some of the limitations of the present study include trying to compare results of pooled genome sequencing (generation 22) to individual mouse sequencing (generation 61: Hillis et al. 2020). Though the alleles of the individual mice can be combined to imitate pooled genome sequences, the differences in number of mice sampled and sampling error make comparisons problematic (see Methods). This is illustrated by the decrease in p-value correlations (between generations 22 and 61) as compared to both the unconstrained and constrained simulations. Nevertheless, as argued above, neither the increase in number of regions detected as differentiated at generation 22 nor the lack of correspondence between detected regions at generations 22 and 61 can be explained solely by methodological differences.

The constraint simulations have their own limitations in that they do not account for male vs female running differences (females run more than males)(Careau et al. 2013). In addition, dominance, epistasis, and gene-environment interactions were not considered. Exclusion of these features may be why we were unable to achieve realistic levels of among-line variation, particularly among the High Runner lines. This model also does not include linkage disequilibrium or realistic rates of recombination.

Additionally we do not include reduction in breeding success across generations, which may explain the drop in selection differential observed by Careau et al. (2013). Lastly, we did not explore the potential effects of relaxing selection for four generations, as when the mice were moved from Wisconsin to California (see Introduction). A cluster of generations of no selection in the HR lines could allow for some drift of the favored alleles.

Although, we are unsure as to why we see so many regions at FDR = 0.01 that do not correspond to the generation 61 findings by Hillis et al. (2020), our simulations suggest that regions with the strongest effect sizes on wheel running are likely to be among the generation 22 regions. Given the statistical significance and number of SNPs identified in our "strictly" culled differentiated regions, these regions are most likely to have had the greatest impact on wheel running at the start of the selection experiment. Among these regions are genes related to olfactory/vomeronasal function, reward pathways, and a miRNA cluster that has been associated with energy homeostasis in neonatal development. All of these associations make sense based on known phenotypic differences between the HR and control lines (see Introduction).

Future directions might include more complex simulations (e.g., see Baldwin-Brown et al. 2014; Stephan 2016; Castro et al. 2019), which may better help to explain the 17X increase in regions detected at generation 22. Including genomic data from more generations (especially from the base population, generations near to but before the selection limit, and current generations [i.e., around 100]) may provide more clarity regarding how the response to selection changes across phases of the selection response (cf. Rose et al. 2005; Castro et al. 2021). Analyses using all loci and a kinship

matrix would enable determination of some interactions between genes. Functional analyses, such as knockouts of some of the genes whose alleles appear to have been favored by selection, may provide direct evidence of influence on wheel-running behavior (e.g., Schmidt et al. 2008; Chaouloff et al. 2011; MacKay et al. 2019). Furthermore, analyses of other physiological aspects of these KO mice may help to better understand the mechanisms by which the gene influences wheel running.

Acknowledgements

We would like to thank Dr. Lei Yu for help with SNP calling, Layla Hiramatsu for assistance in sample collection, and Dr. Zhenyu (Arthur) Jia, Dr. Shizhong Xu, Dr. Frank Chan, and Dr. Tony Long for comments and suggestions for this study.

Competing of interests

The authors have no competing interests.

Author Contributions

Conceptualization, D.A.H., Lir.Y., F.P.M.dEV., D.P., T.G.; investigation, D.A.H., Lir.Y., G.M.W., F.P.M.dEV., D.P., T.G.; software, D.A.H., Lei.Y.; formal analysis, D.A.H., Lir.Y., T.G.; writing – original draft, D.A.H., T.G.; writing review and editing, all authors.

Data Availability

Generation 61 data were made available by Hillis et al. (2020) and can be found at https://doi.org/10.25386/genetics.12436649. Generation 22 fastq files are available on the SRA database, accession = PRJNA758905 (https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA758905).

References

Ahrens WH, Cox DJ, Budhwar G. 1990. Use of the arcsine and square root transformations for subjectively determined percentage data. *Weed Sci.* 38:452–458.

Al-Murrani WK, Roberts RC. 1974. Genetic variation in a line of mice selected to its limit for high body weight. *Anim. Sci.* 19:273–289.

Baldi P, Long AD. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17:509–519.

Baldwin-Brown JG, Long AD, Thornton KR. 2014. The power to detect quantitative trait loci using resequenced, experimentally evolved populations of diploid, sexual organisms. *Mol. Biol. Evol.* 31:1040–1055.

Bamforth SD, Bragança J, Farthing CR, Schneider JE, Broadbent C, Michell AC, Clarke K, Neubauer S, Norris D, Brown NA, et al. 2004. Cited2 controls left-right patterning and heart development through a Nodal-Pitx2c pathway. *Nat. Genet.* 36:1189–1196.

Barbera JPM, Rodrigues TA, Greene NDE, Weninger WJ, Simeone A, Copp AJ, Beddington RSP, Dunwoodie S. 2002. Folic acid prevents exencephaly in Cited2 deficient mice. *Hum. Mol. Genet.* 11:283–293.

Barton NH, Turelli M. 1989. Evolutionary quantitative genetics: how little do we know? *Annu. Rev. Genet.* 23:337–370.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57:289–300.

Brown WP, Bell AE. 1961. Genetic analysis of a “plateaued” population of *Drosophila melanogasteri*. *Genetics* 46:407–425.

Bult A, Lynch CB. 2000. Breaking through artificial selection limits of an adaptive behavior in mice and the consequences for correlated responses. *Behav. Genet.* 30:193–206.

Burke MK, Dunham JP, Shahrestani P, Thornton KR, Rose MR, Long AD. 2010. Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature* 467:587–590.

Cadney MD, Hiramatsu L, Thompson Z, Zhao M, Kay JC, Singleton JM, Albuquerque RL de, Schmill MP, Saltzman W, Garland, Jr. T. 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. *Physiol. Behav.* 234:113389.

Cardinali B, Cappella M, Provenzano C, Garcia-Manteiga JM, Lazarevic D, Cittaro D, Martelli F, Falcone G. 2016. MicroRNA-222 regulates muscle alternative splicing through Rbm24 during differentiation of skeletal muscle cells. *Cell Death Dis.* 7:e2086–e2086.

Careau V, Wolak ME, Carter PA, Garland, Jr. T. 2013. Limits to behavioral evolution: the quantative genetics of a complex trait under directional selection. *Evolution* 67:3102–3119.

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

Castro AA, Garland, Jr. T, Ahmed S, Holt NC. 2022. Trade-offs in muscle physiology in selectively bred High Runner mice. *J. Exp. Biol.*:jeb.244083.

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

Castro JP, Yancoskie MN, Marchini M, Belohlavy S, Hiramatsu L, Kucka M, Beluch NH, Rolian C, Chan YF. 2019. An integrative genomic analysis of the Longshanks selection experiment for longer limbs in mice. *eLife* 8:e40214.

Chaouloff F, Dubreucq S, Bellocchio L, Marsicano G. 2011. Endocannabinoids and motor behavior: CB1 receptors also control running activity. *Physiology* 26:76–77.

Copes LE, Schutz H, Dlugsoz EM, Judex S, Garland, Jr. T. 2018. Locomotor activity, growth hormones, and systemic robusticity: An investigation of cranial vault thickness in mouse lines bred for high endurance running. *Am. J. Phys. Anthropol.* 166:442–458.

Dewan I, Garland, Jr. T, Hiramatsu L, Careau V. 2019. I smell a mouse: indirect genetic effects on voluntary wheel-running distance, duration and speed. *Behav. Genet.* 49:49–59.

Didion JP, Morgan AP, Yadgary L, Bell TA, McMullan RC, Ortiz de Solorzano L, Britton-Davidian J, Bult CJ, Campbell KJ, Castiglia R, et al. 2016. *R2d2* drives selfish sweeps in the house mouse. *Mol. Biol. Evol.* 33:1381–1395.

Dlugosz EM, Schutz H, Meek TH, Acosta W, Downs CJ, Platzer EG, Chappell MA, Garland, Jr. T. 2013. Immune response to a Trichinella spiralis infection in house mice from lines selectively bred for high voluntary wheel running. *J. Exp. Biol.* 216:4212–4221.

1220 Dobzhansky T, Spassky B. 1969. Artificial and natural selection for two behavioral traits
1221 in *Drosophila pseudoobscura*. *Proc. Natl. Acad. Sci. U. S. A.* 62:75–80.

1222 Douhard F, Douhard M, Gilbert H, Monget P, Gaillard J, Lemaître J. 2021. How much
1223 energetic trade-offs limit selection? Insights from livestock and related laboratory
1224 model species. *Evol. Appl.* 14:2726–2749.

1225 Dumke CL, Rhodes JS, Garland, Jr. T, Maslowski E, Swallow JG, Wetter AC, Cartee
1226 GD. 2001. Genetic selection of mice for high voluntary wheel running: effect on
1227 skeletal muscle glucose uptake. *J. Appl. Physiol.* 91:1289–1297.

1228 Falconer DS. 1989. Introduction to quantitative genetics. 3rd ed. Burnt Mill, Harlow,
1229 Essex, England : New York: Longman, Scientific & Technical ; Wiley

1230 Garland, Jr. T, Freeman PW. 2005. Selective breeding for high endurance running
1231 increases hindlimb symmetry. *Evolution* 59:1851–1854.

1232 Garland, Jr. T, Kelly SA, Malisch JL, Kolb EM, Hannon RM, Keeney BK, Van Cleave
1233 SL, Middleton KM. 2011. How to run far: multiple solutions and sex-specific
1234 responses to selective breeding for high voluntary activity levels. *Proc. R. Soc. B*
1235 *Biol. Sci.* 278:574–581.

1236 Garland, Jr. T, Schutz H, Chappell MA, Keeney BK, Meek TH, Copes LE, Acosta W,
1237 Drenowatz C, Maciel RC, van Dijk G, et al. 2011. The biological control of
1238 voluntary exercise, spontaneous physical activity and daily energy expenditure in
1239 relation to obesity: human and rodent perspectives. *J. Exp. Biol.* 214:206–229.

1240 Garland, Jr. T, Zhao M, Saltzman W. 2016. Hormones and the evolution of complex
1241 traits: insights from artificial selection on behavior. *Integr. Comp. Biol.* 56:207–
1242 224.

1243 Garland T, Downs CJ, Ives AR. 2022. Trade-offs (and constraints) in organismal
1244 biology. *Physiol. Biochem. Zool.* 95:82–112.

1245 Grifone R, Saquet A, Desgres M, Sangiorgi C, Gargano C, Li Z, Coletti D, Shi D-L.
1246 2021. Rbm24 displays dynamic functions required for myogenic differentiation
1247 during muscle regeneration. *Sci. Rep.* 11:9423.

1248 Guderley H, Joanisse DR, Mokas S, Bilodeau GM, Garland, Jr. T. 2008. Altered fibre
1249 types in gastrocnemius muscle of high wheel-running selected mice with mini-
1250 muscle phenotypes. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* 149:490–
1251 500.

1252 Hillis DA, Garland, Jr. T. 2022. Multiple solutions at the genomic level in response to
1253 selective breeding for high locomotor activity. *Genetics*.

1254 Hillis DA, Yadgary L, Weinstock GM, Pardo-Manuel de Villena F, Pomp D, Fowler AS,
1255 Xu S, Chan F, Garland, Jr. T. 2020. Genetic basis of aerobically supported

voluntary exercise: results from a selection experiment with house mice. *Genetics* 216:781–804.

Hiramatsu L, Kay JC, Thompson Z, Singleton JM, Claghorn GC, Albuquerque RL, Ho Brittany, Ho Brett, Sanchez G, Garland, Jr. T. 2017. Maternal exposure to Western diet affects adult body composition and voluntary wheel running in a genotype-specific manner in mice. *Physiol. Behav.* 179:235–245.

Jani PH, Gibson MP, Liu C, Zhang H, Wang X, Lu Y, Qin C. 2016. Transgenic expression of Dspp partially rescued the long bone defects of Dmp1-null mice. *Matrix Biol.* 52–54:95–112.

Kay JC, Colbath J, Talmadge RJ, Garland, Jr. T. 2022. Mice from lines selectively bred for voluntary exercise are not more resistant to muscle injury caused by either contusion or wheel running.Kubis H-P, editor. *PLOS ONE* 17:e0278186.

Kelly SA, Czech PP, Wight JT, Blank KM, Garland, Jr. T. 2006. Experimental evolution and phenotypic plasticity of hindlimb bones in high-activity house mice. *J. Morphol.* 267:360–374.

Kelly SA, Gomes FR, Kolb EM, Malisch JL, Garland, Jr. T. 2017. Effects of activity, genetic selection and their interaction on muscle metabolic capacities and organ masses in mice. *J. Exp. Biol.* 220:1038–1047.

Khan RH, Rhodes JS, Girard IA, Schwartz NE, Garland, Jr. T. 2024. Does behavior evolve first? Correlated responses to artificial selection for voluntary wheel-running behavior in house mice. *Ecol. Evol. Physiol.* 97.

Kolb EM, Kelly SA, Garland, Jr. T. 2013. Mice from lines selectively bred for high voluntary wheel running exhibit lower blood pressure during withdrawal from wheel access. *Physiol. Behav.* 112–113:49–55.

Kolb EM, Kelly SA, Middleton KM, Sermsakdi LS, Chappell MA, Garland, Jr. T. 2010. Erythropoietin elevates VO_2,max but not voluntary wheel running in mice. *J. Exp. Biol.* 213:510–519.

Kolb EM, Rezende EL, Holness L, Radtke A, Lee SK, Obenaus A, Garland, Jr. T. 2013. Mice selectively bred for high voluntary wheel running have larger midbrains: support for the mosaic model of brain evolution. *J. Exp. Biol.* 216:515–523.

Lerner IM, Dempster ER. 1951. Attenuation of genetic progress under continued selection in poultry. *Heredity* 5:75–94.

Lightfoot JT, De Geus EJC, Booth FW, Bray MS, Den Hoed M, Kaprio J, Kelly SA, Pomp D, Saul MC, Thomis MA, et al. 2018. Biological/genetic regulation of physical activity level: Consensus from GenBioPAC. *Med. Sci. Sports Exerc.* 50:863–873.

1292 Lillie M, Honaker CF, Siegel PB, Carlborg Ö. 2019. Bidirectional selection for body
1293 weight on standing genetic variation in a chicken model.
1294 *Genes*|*Genomes*|*Genetics*:g3.400038.2019.

1295 Long A, Liti G, Luptak A, Tenaillon O. 2015. Elucidating the molecular architecture of
1296 adaptation via evolve and resequence experiments. *Nat. Rev. Genet.* 16:567–
1297 582.

1298 MacDonald ST, Bamforth SD, Chen C-M, Farthing CR, Franklyn A, Broadbent C,
1299 Schneider JE, Saga Y, Lewandoski M, Bhattacharya S. 2008. Epiblastic Cited2
1300 deficiency results in cardiac phenotypic heterogeneity and provides a mechanism
1301 for haploinsufficiency. *Cardiovasc. Res.* 79:448–457.

1302 MacKay H, Scott CA, Duryea JD, Baker MS, Laritsky E, Elson AE, Garland, Jr. T,
1303 Fiorotto ML, Chen R, Li Y, et al. 2019. DNA methylation in AgRP neurons
1304 regulates voluntary exercise behavior in mice. *Nat. Commun.* [Internet] 10.
1305 Available from: http://www.nature.com/articles/s41467-019-13339-3

1306 Malisch JL, Saltzman W, Gomes FR, Rezende EL, Jeske DR, Garland, Jr. T. 2007.
1307 Baseline and stress-induced plasma corticosterone concentrations of mice
1308 selectively bred for high voluntary wheel running. *Physiol. Biochem. Zool.*
1309 80:146–156.

1310 Mathes WF, Nehrenberg DL, Gordon R, Hua K, Garland, Jr. T, Pomp D. 2010.
1311 Dopaminergic dysregulation in mice selectively bred for excessive exercise or
1312 obesity. *Behav. Brain Res.* 210:155–163.

1313 Meek TH, Lonquich BP, Hannon RM, Garland, Jr. T. 2009. Endurance capacity of mice
1314 selectively bred for high voluntary wheel running. *J. Exp. Biol.* 212:2908–2917.

1315 Middleton KM, Goldstein BD, Guduru PR, Waters JF, Kelly SA, Swartz SM, Garland, Jr.
1316 T. 2010. Variation in within-bone stiffness measured by nanoindentation in mice
1317 bred for high levels of voluntary wheel running. *J. Anat.* 216:121–131.

1318 Middleton KM, Shubin CE, Moore DC, Carter PA, Garland, Jr. T, Swartz SM. 2008. The
1319 relative importance of genetics and phenotypic plasticity in dictating bone
1320 morphology and mechanics in aged mice: Evidence from an artificial selection
1321 experiment. *Zoology* 111:135–147.

1322 Nguyen QAT, Hillis D, Katada S, Harris T, Pontrello C, Garland, Jr. T, Haga-Yamanaka
1323 S. 2020. Coadaptation of the chemosensory system with voluntary exercise
1324 behavior in mice. *PLOS ONE* 15:e0241758.

1325 Porto A, Schmelter R, VandeBerg JL, Marroig G, Cheverud JM. 2016. Evolution of the
1326 genotype-to-phenotype map and the cost of pleiotropy in mammals. *Genetics*
1327 204:1601–1612.

1328 Reeve JP. 2000. Predicting long-term response to selection. *Genet. Res.* 75:83–94.

1329 Reeve JP, Fairbairn DJ. 2001. Predicting the evolution of sexual size dimorphism:
1330 Predicting the evolution of SSD. *J. Evol. Biol.* 14:244–254.

1331 Rhodes JS, Gammie SC, Garland, Jr. T. 2005. Neurobiology of mice selected for high
1332 voluntary wheel-running activity. *Integr. Comp. Biol.* 45:438–455.

1333 Rhodes JS, Hosack GR, Girard I, Kelley AE, Mitchell GS, Garland, Jr. T. 2001.
1334 Differential sensitivity to acute administration of cocaine, GBR 12909, and
1335 fluoxetine in mice selectively bred for hyperactive wheel-running behavior.
1336 *Psychopharmacology (Berl.)* 158:120–131.

1337 Rhodes JS, van Praag H, Jeffrey S, Girard I, Mitchell GS, Garland, Jr. T, Gage FH.
1338 2003. Exercise increases hippocampal neurogenesis to high levels but does not
1339 improve spatial learning in mice bred for increased voluntary wheel running.
1340 *Behav. Neurosci.* 117:1006–1016.

1341 Roberts RC. 1966. The limits to artificial selection for body weight in the mouse II. The
1342 genetic nature of the limits. *Genet. Res.* 8:361–375.

1343 Rose MR, Passananti HB, Chippindale AK, Phelan JP, Matos M, Teotónio H, Mueller
1344 LD. 2005. The effects of evolution are local: evidence from experimental
1345 evolution in drosophila. *Integr. Comp. Biol.* 45:486–491.

1346 Schlötterer C, Kofler R, Versace E, Tobler R, Franssen SU. 2015. Combining
1347 experimental evolution with next-generation sequencing: a powerful tool to study
1348 adaptation from standing genetic variation. *Heredity* 114:431–440.

1349 Schmidt S, Gawlik V, Hölter SM, Augustin R, Scheepers A, Behrens M, Wurst W,
1350 Gailus-Durner V, Fuchs H, de Angelis MH, et al. 2008. Deletion of glucose
1351 transporter GLUT8 in mice increases locomotor activity. *Behav. Genet.* 38:396–
1352 406.

1353 Schwartz NL, Patel BA, Garland, Jr. T, Horner AM. 2018. Effects of selective breeding
1354 for high voluntary wheel-running behavior on femoral nutrient canal size and
1355 abundance in house mice. *J. Anat.* 233:193–203.

1356 Sella G, Barton NH. 2019. Thinking about the evolution of complex traits in the era of
1357 genome-wide association studies. *Annu. Rev. Genomics Hum. Genet.* 20:461–
1358 493.

1359 Stephan W. 2016. Signatures of positive selection: from selective sweeps at individual
1360 loci to subtle allele frequency changes in polygenic adaptation. *Mol. Ecol.* 25:79–
1361 88.

1362 Stratford T, Logan C, Zile M, Maden M. 1999. Abnormal anteroposterior and
1363 dorsoventral patterning of the limb bud in the absence of retinoids. *Mech. Dev.*
1364 81:115–125.

1365 Swallow JG, Carter PA, Garland, Jr. T. 1998. Artificial selection for increased wheel-
1366 running behavior in house mice. *Behav. Genet.* 28:227–237.

1367 Swallow JG, Garland, Jr. T, Carter PA, Zhan W-Z, Sieck GC. 1998. Effects of voluntary
1368 activity and genetic selection on aerobic capacity in house mice (*Mus*
1369 *domesticus*). *J. Appl. Physiol.* 84:69–76.

1370 Swallow JG, Hayes JP, Koteja P, Garland, Jr. T. 2009. Selection experiments and
1371 experimental evolution of performance and physiology. In: Experimental
1372 evolution: concepts, methods, and applications of selection experiments.
1373 Berkeley: University of California Press. p. 301–351.

1374 Syme DA, Evashuk K, Grintuch B, Rezende EL, Garland, Jr. T. 2005. Contractile
1375 abilities of normal and “mini” triceps surae muscles from mice (*Mus domesticus*)
1376 selectively bred for high voluntary wheel running. *J. Appl. Physiol.* 99:1308–1316.

1377 Thompson Z, Argueta D, Garland, Jr. T, DiPatrizio N. 2017. Circulating levels of
1378 endocannabinoids respond acutely to voluntary exercise, are altered in mice
1379 selectively bred for high voluntary wheel running, and differ between the sexes.
1380 *Physiol. Behav.* 170:141–150.

1381 Travisano M, Shaw RG. 2013. Lost in the map. *Evolution* 67:305–314.

1382 Val P, Martinez-Barbera J-P, Swain A. 2007. Adrenal development is initiated by Cited2
1383 and Wt1 through modulation of Sf-1 dosage. *Development* 134:2349–2358.

1384 Verdelis K, Ling Y, Sreenath T, Haruyama N, MacDougall M, Van Der Meulen MCH,
1385 Lukashova L, Spevak L, Kulkarni AB, Boskey AL. 2008. DSPP effects on in vivo
1386 bone mineralization. *Bone* 43:983–990.

1387 Wallace IJ, Garland, Jr. T. 2016. Mobility as an emergent property of biological
1388 organization: Insights from experimental evolution: Mobility and biological
1389 organization. *Evol. Anthropol. Issues News Rev.* 25:98–104.

1390 Wallace IJ, Middleton KM, Lublinsky S, Kelly SA, Judex S, Garland, Jr. T, Demes B.
1391 2010. Functional significance of genetic variation underlying limb bone
1392 diaphyseal structure. *Am. J. Phys. Anthropol.* 143:21–30.

1393 Wallace IJ, Tommasini SM, Judex S, Garland T, Demes B. 2012. Genetic variations and
1394 physical activity as determinants of limb bone morphology: An experimental
1395 approach using a mouse model. *Am. J. Phys. Anthropol.* 148:24–35.

1396 Wang Z, Emmerich A, Pillon NJ, Moore T, Hemerich D, Cornelis MC, Mazzaferro E,
1397 Broos S, Ahluwalia TS, Bartz TM, et al. 2022. Genome-wide association
1398 analyses of physical activity and sedentary behavior provide insights into
1399 underlying mechanisms and roles in disease prevention. *Nat. Genet.* 54:1332–
1400 1344.

Waters RP, Pringle RB, Forster GL, Renner KJ, Malisch JL, Garland, Jr. T, Swallow JG. 2013. Selection for increased voluntary wheel-running affects behavior and brain monoamines in mice. *Brain Res.* 1508:9–22.

Wood AR, The Electronic Medical Records and Genomics (eMERGE) Consortium, The MIGen Consortium, The PAGE Consortium, The LifeLines Cohort Study, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* 46:1173–1186.

Xie Y, Pan W, Khodursky AB. 2005. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics* 21:4280–4288.

Xu S, Garland T. 2017. A mixed model approach to genome-wide association studies for selection signatures, with application to mice bred for voluntary exercise behavior. *Genetics* 207:785–799.

Zamer WE, Scheiner SM. 2014. A conceptual framework for organismal biology: linking theories, models, and data. *Integr. Comp. Biol.* 54:736–756.

Zhang M, Han Y, Liu J, Liu L, Zheng L, Chen Y, Xia R, Yao D, Cai X, Xu X. 2020. Rbm24 modulates adult skeletal muscle regeneration via regulation of alternative splicing. *Theranostics* 10:11159–11177.

Zinski AL, Carrion S, Michal JJ, Gartstein MA, Quock RM, Davis JF, Jiang Z. 2021. Genome-to-phenome research in rats: progress and perspectives. *Int. J. Biol. Sci.* 17:119–133.

Supporting Information

S1 Fig. Simulation power results by effect size and generation. Power (Y-axis) of different effect sizes at different generations (X-axis). Effect Size - Color: 204.8 - brown, 102.4 - red, 51.2 - orange, 25.6 - yellow, 12.8 - dark green, 6.4 - light green, 3.2 - dark blue, 1.6 - light blue, 0.8 - dark purple, 0.4 - light purple

S1 File. Regularized and windowed regularized F-test (WRT). Description of methodology and rationale.

S2 File. Parameters and seeds for constraint simulations. List of the parameters and seeds used for the simulations with and without constraints.

S1 Table. Effect size and Type I error rates. Includes error rates and means for different effect sizes and sample sizes.

S2 Table. Power to detect differentiation between HR and C lines in relation to effect size of locus and generation under two simulation models. Includes effect sizes, sample sizes, revolutions when homozygous, and power and mean for each generation and model.

S3 Table. Differentiated regions identified at generation 22 (FDR = 0.01). Includes chromosomal location, size of region, most statistically significant base pair p-value, and position of this base pair.

S4 Table. Genes included in "strict" culling regions at generation 22.

S5 Table. Gene ontology results for generation 22 "strict" culling genes. Includes GO terms, fold enrichment, and raw p-values.