



Causes of variability in estimates of mutational variance from mutation accumulation experiments

Cara Conradsen , * Mark W. Blows , Katrina McGuigan 

School of Biological Sciences, The University of Queensland, St. Lucia, QLD 4072, Australia

*Corresponding author: School of Biological Sciences, The University of Queensland, St. Lucia, QLD 4072 Australia. Email: cara.conradsen@uqconnect.edu.au

Abstract

Characteristics of the new phenotypic variation introduced via mutation have broad implications in evolutionary and medical genetics. Standardized estimates of this mutational variance, V_M , span 2 orders of magnitude, but the causes of this remain poorly resolved. We investigated estimate heterogeneity using 2 approaches. First, meta-analyses of ~150 estimates of standardized V_M from 37 mutation accumulation studies did not support a difference among taxa (which differ in mutation rate) but provided equivocal support for differences among trait types (life history vs morphology, predicted to differ in mutation rate). Notably, several experimental factors were confounded with taxon and trait, and further empirical data are required to resolve their influences. Second, we analyzed morphological data from an experiment in *Drosophila serrata* to determine the potential for unintentional heterogeneity among environments in which phenotypes were measured (i.e. among laboratories or time points) or transient segregation of mutations within mutation accumulation lines to affect standardized V_M . Approximating the size of an average mutation accumulation experiment, variability among repeated estimates of (accumulated) mutational variance was comparable to variation among published estimates of standardized V_M . This heterogeneity was (partially) attributable to unintended environmental variation or within line segregation of mutations only for wing size, not wing shape traits. We conclude that sampling error contributed substantial variation within this experiment, and infer that it will also contribute substantially to differences among published estimates. We suggest a logistically permissive approach to improve the precision of estimates, and consequently our understanding of the dynamics of mutational variance of quantitative traits.

Keywords: meta-analysis; *Drosophila*; life history; morphology; heritability; coefficient of variance; wing; sampling error

Introduction

The magnitude of per-generation increase in genetic variance due to spontaneous mutations (V_M) is important for a wide range of genetic and evolutionary phenomena, including the maintenance of quantitative genetic variance (Lynch 1988; Barton and Turelli 1989; Johnson and Barton 2005). Much of our understanding of V_M comes from mutation accumulation (MA) experiments, where populations diverge phenotypically due solely to the neutral fixation of new mutations (Mukai 1964; Halligan and Keightley 2009). Reviews of MA experiments in a range of traits and taxa have reported that mutation increases phenotypic variance in quantitative traits by 10^{-4} – 10^{-2} times the environmental variance of the trait, or 0.02–5.1% of the trait mean per generation (Houle et al. 1996; Lynch et al. 1999; Halligan and Keightley 2009). Differences in V_M may cause differences in the magnitude of standing quantitative genetic variation and, ultimately, in rates of phenotypic evolution (Houle 1998; Lynch et al. 1999; Houle et al. 2017; Walsh and Lynch 2018). However, the causes of variation among estimates of V_M , and thus the evolutionary interpretation of this variability, are not well resolved.

Mutation rate is known to vary widely among species (reviewed in Katju and Bergthorsson 2019), with further opportunity for differences in per-generation mutation number arising

through differences in ploidy, genome size, and/or effective population size (Lynch et al. 1999; Lynch 2010; Sung et al. 2012). Marked variation in mutation rate has also been observed within species, both among replicated MA experiments (i.e. different founder genotypes: Ness et al. 2012; Sung et al. 2012; Schrider et al. 2013; Ho et al. 2020) and among lines within a single MA panel (Huang et al. 2016; Ho et al. 2020). Resulting differences in mutation number may explain variation in V_M estimates, such as, for example, the 4-times difference in h_M^2 of body size estimated for different MA in *Caenorhabditis elegans* (Azevedo et al. 2002; Estes et al. 2005; Ostrow et al. 2007).

Traits have also been hypothesized to differ in magnitude of V_M due to differences in mutation rate, arising due to differences in the number of contributing loci. Specifically, life history traits are hypothesized to be affected by more loci than morphological traits (Houle 1991, 1992, 1998; Houle et al. 1996; Merilä and Sheldon 1999). The magnitude of V_M depends not only on the rate of mutation, but also on their effects, and the relationship between rate and effect size is not well characterized. Besnard et al. (2020) demonstrated that the high mutational variance (and relatively rapid evolution) of a vulval phenotype in nematodes was due to a broad mutational target size, rather than large-effect mutation. Whether trait types differ systematically in mutational target size is difficult to assess, as a full catalog of causal loci is

Received: February 25, 2022. Accepted: April 08, 2022

© The Author(s) 2022. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

unknown for most quantitative traits (Barton and Keightley 2002; Mackay et al. 2009; Yang et al. 2010; Rockman 2012). Indeed, emerging evidence that diverse traits, including morphology, are all highly polygenic (Yang et al. 2010; Boyle et al. 2017) suggests that differences in the distribution of mutational effect sizes (Simons et al. 2018), rather than simply in number of contributing loci, might cause heterogeneity in estimates of V_M among traits.

Comparison among trait types is complicated by differences among them in variability and measurement scale, which may influence standardized values. Low mutational heritability ($h_M^2 = V_M/V_E$, where V_E is the environmental variance) of life-history traits relative to morphological traits has been attributed to greater environmental variance (larger V_E) for life-history traits, rather than lower V_M (Houle et al. 1996). Thus, comparison on the coefficient of mutational variance scale (CV_M ; $100 \times \sqrt{V_M}/\bar{X}$, where \bar{X} is the trait mean) reveals a different picture, one of greater mutational variance in life history than morphological traits, consistent with the prediction of greater mutational target size (Houle et al. 1996).

Other contributions to variation in magnitude of V_M might be revealed by consideration of the MA experimental design itself. The timeframe over which mutations accumulate might influence estimates of V_M . When MA lines are established from a homozygous (heterozygous) base population, estimates of V_M will be downwardly (upwardly) biased before $6N_e$ generations (Lynch and Hill 1986). However, V_M is typically estimated after $> 6N_e$ generations, suggesting limited contribution of ancestral variation to variability of V_M . Conversely, long-running MA experiments might under-estimate V_M when the cumulative effect of low fitness mutation causes line extinction, or within-line selection against further accumulation (Lynch et al. 1999; Estes et al. 2004; McGuigan and Blows 2013). A decline in V_M over time has been observed in some studies (Mackay et al. 1995), but not in others (García-Dorado et al. 2000; Hall et al. 2008).

Stochastic sampling from the distribution of mutational effects could also introduce temporal heterogeneity among estimates of V_M . For example, among-line variance estimated before vs after a line(s) fixed a large effect mutation(s) could result in inference of a much larger per-generation increase in variance at the second time-point relative to the first. Transient within-line segregation of mutations might generate variability in estimates, for example causing temporary inflation of within-line variance (V_E), impacting power to detect among-line variance, and potentially biasing estimates of h_M^2 (downward) and CV_M (upward; see Hoffmann et al. 2016). Notably, several studies in nematode have suggested that within-line variance increased over the duration of the MA experiment (Baer 2008; Baer et al. 2010; Braendle et al. 2010), which may contribute to a pattern of lower estimated V_M in longer-running MA experiments.

Environmental context within which MA lines are assayed could also contribute to variation among V_M estimates. Several studies have considered the effect of replicable, experimenter-imposed, changes in the environment, including in temperature (Wayne and Mackay 1998), light (Kavanaugh and Shaw 2005), and density (Fry et al. 1996). Although the magnitude of V_M often varies under such environmental manipulations, there is only weak evidence for predictable patterns, such as novel or stressful environments increasing the magnitude of V_M (Kondrashov and Houle 1994; Martin and Lenormand 2006). Even in carefully controlled laboratory experiments, factors such as food quality or quantity, light, humidity and diurnal timing of collection will vary among individuals or lines within a phenotyping assay, and

among assays conducted in different laboratories or at different times within the same laboratory. Such variation may impact estimates of standardized V_M through inflation of within-line variance, similar to the effect of transient, within-line segregation of new mutations. Furthermore, if MA lines differ in their response to this unintended environmental heterogeneity, then genotype by environment ($G \times E$) variance could contribute variation among MA lines, and variability among estimates of V_M , a potential source of variation that has received little attention (but see García-Dorado et al. 2000).

Here, we combined 2 approaches to investigate causes of variability in estimates of mutational variance. Given that it has been over 20 years since this variability has been broadly documented and investigated (Houle et al. 1996; Houle 1998; Lynch et al. 1999), we first conducted a meta-analysis to update tests of the previously implicated causal factors of taxon (Lynch and Walsh 1998; Lynch et al. 1999; Halligan and Keightley 2009) and trait type (Houle et al. 1996; Houle 1998). We had intended to examine how the number of generations affected V_M (Lynch and Hill 1986; Mackay et al. 1995), but MA duration was confounded with taxon (detailed in the Results). Second, we conducted a new empirical experiment in *Drosophila serrata*, in which we repeatedly estimate the among-line (mutational) variance to investigate whether unintended environmental heterogeneity, or transient within-line segregation of mutations can contribute variation among estimates. After accounting for these effects within the data, we finally quantify the magnitude of variation among estimates from a set of 10 wing shape traits to characterize the magnitude of variation among estimates within a trait category.

Methods

Meta-analysis of empirical estimates of mutational variance

Literature search

We extracted all studies in 7 reviews of mutational variance: Lynch (1988), Keightley et al. (1993), Houle et al. (1996), Lynch and Walsh (1998), Lynch et al. (1999), Halligan and Keightley (2009), and Walsh and Lynch (2018). We then searched the Web of Science database on 11/12/2019 at 4:38 p.m. AEST for journal article document types meeting the topic criteria of “MA” and (varia* or “mutat* coefficient”) and published between 1998 and 2019. These years overlapped Halligan and Keightley (2009) (fitness traits only) and Walsh and Lynch (2018) (brief update on Lynch and Walsh 1998), allowing us to capture papers that may have been excluded from those reviews, as well as those published subsequently.

Further details on the papers identified and preliminary handling steps can be found in Supplementary Fig. 1. For 473 unique papers identified, we screened titles and abstracts, then the full text, for relevance, applying 4 strict criteria, retaining only studies where the estimates of mutational variance were: (1) quantitative; (2) from spontaneous MA; (3) from MA environmental conditions; and (4) not re-reporting of previously published estimates. We excluded 6 studies of transcriptomic data as the number of traits was much larger than for other trait categories.

Meta-analysis data collection

For each of the 65 papers retained after applying the above criteria, mutational parameter estimates were extracted (as described in Supplementary Table 1), associated with taxon and trait identifiers, and details of the experimental design. Twenty papers not

reporting error for the mutational parameters were excluded (Supplementary Table 1c). Following initial qualitative assessments of data, we excluded 5 studies (15 traits) due to low representation of taxon type (one vertebrate, *Mus musculus*; 1 alga, *Chlamydomonas reinhardtii*, and 2 non-*Drosophila* insects: *Daktulosphaira vitifoliae* and *Nasonia vitripennis*), and 1 study due to low representation of trait type: mitotic cell division traits (Supplementary Table 1b).

Where possible, we extracted (or calculated from provided information) both the coefficient of variance (CV_M ; $100 \times \sqrt{V_M}/\bar{X}$, where \bar{X} is the trait mean) and mutational heritability (h_M^2 ; V_M/V_E , where V_E is the environmental variance) for each trait. As detailed below, estimates were weighted by the inverse of their standard error (SE) in the meta-analysis. Where these were not reported for h_M^2 or CV_M , but were for V_M and V_E or \bar{X} , we used a sampling approach to obtain estimates. We sampled from $N \sim (\hat{\theta}, V)$ 10,000 times, using the *morm* function in R [v. 3.6.1], where $\hat{\theta}$ and V were respectively the reported parameter value and its SE. We then calculated CV_M or h_M^2 for each of these simulated samples, and obtained the SE of this sample of estimates. Samples with negative values of V_M are undefined for CV_M ; to ensure unbiased estimates of the magnitude of error we calculated CV_M as: $100 \times (\text{sign of } V_M) \times \sqrt{|V_M|}/\bar{X}$. This sampling approach was used to estimate the error for 28% of the h_M^2 estimates and 61% of the CV_M estimates analyzed (Supplementary Table 1a). Two studies (17 estimates) were excluded due to nonsensically large SE estimates, while a further 3 estimates (from 3 studies) were excluded due to nonsensical scaled parameter estimates (detailed in Supplementary Table 1b). Two extreme values (>3 SD) of h_M^2 and 2 of CV_M were excluded from the analyses (Supplementary Table 1b). There were 11 cases with extremely small SE (>5 IQR below the median); notably, 6 of these came from studies where confidence intervals (CIs) were constrained to be positive, suggesting that this boundary condition had reduced the SE estimate, inflating meta-analysis weights for traits where the mutational variance was not supported. These outliers were excluded from analyses, although results and conclusions were qualitatively consistent when they were included.

Predictor variables for the meta-analysis

Estimates came from 11 species, and based on the distribution of estimates, we defined 5 taxon categories (Fig. 1a): *Daphnia* (*Daphnia pulex* only); *Drosophila* (*Drosophila melanogaster* [$n=68$] and *D. serrata* [$n=5$]); Plant (*Arabidopsis thaliana* [$n=12$], *Amsinckia douglasiana* [$n=2$] and *Amsinckia gloriosa* [$n=1$]) and; Nematode (*C. elegans* [$n=62$], *C. brenneri* [$n=4$], *C. briggsae* [$n=8$], *C. remanei* [$n=5$], and *Oscheius myriophila* [$n=5$]). We differ from a previous

study Houle et al. (1996, 1998) in considering size of juveniles as morphological (not growth) traits. Reflecting more recent publications, we defined a physiology category (33% of estimates; Fig. 1b), which included locomotive, enzymatic and metabolic activity traits (Supplementary Table 2), which may differ from life-history or morphological traits in mutational target size or environmental sensitivity. We assigned the relatively well-represented life-history traits (52% of estimates; Fig. 1b) into more narrowly defined subcategories: total fitness, survival, productivity, and a miscellaneous category (capturing traits such as development time, phenology, longevity and mating success, which were individually less well represented) (Fig. 1b; Supplementary Table 2).

Meta-analyses of mutational variance estimates

We implemented a mixed model analyses via PROC MIXED in SAS v.9.4 (SAS Institute Inc., Cary, N.C.), using restricted-maximum likelihood (REML) and applying the Satterthwaite approximation to correct the denominator degrees of freedom, to fit the model:

$$y_{ijkl} = \mu + \text{Taxon}_i + \text{Trait}_j + \text{Study}_k + \epsilon_{ijkl}, \quad (1)$$

where y was the vector of published estimates (either h_M^2 or CV_M), and μ was the grand mean; the categorical predictors of taxon and trait (defined above) were fit as fixed effects. Estimates were weighted by the inverse of the SE of h_M^2 or CV_M , obtained as detailed above. The study was fit as a random effect, accounting for nonindependence among estimates within a paper (1–29 estimates per study; median=3). Studies reporting multiple estimates varied widely in whether these were estimates from different trait types, species (strains), sexes, or time points. Likely reflecting this, most variation not accounted for by the fixed effects was observed at the residual, not study, level (99% for h_M^2 ; 85% for CV_M). Similarly, while some studies shared the same MA lines (Supplementary Table 3) fitting a further random effect to account for this nonindependence did not explain any variation, a likely consequence of both the unbalanced design (only some studies share lines), and the relative variation of estimates. We investigated different options for fitting heterogeneous residuals (e.g. allowing separate estimation of residuals for studies grouped depending on the number of estimates reported), but interpretation of the fixed effects (taxon and trait) were consistent across all investigated models, and we report results only from model (1) above.

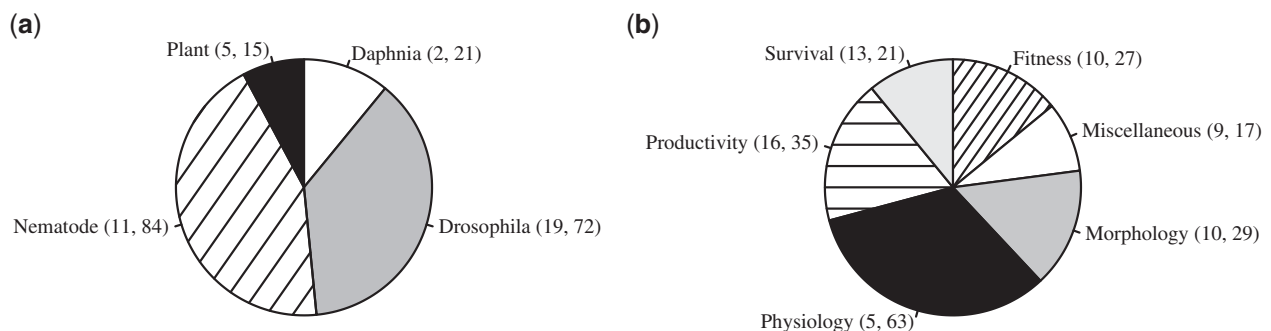


Fig. 1. The distribution of published estimates of mutational variance across taxon (a) and trait (b) categories. The number of studies (first value in brackets) and estimates (second value in brackets) per category are shown. See Methods (and Supplementary Table 2) for details of the categories and Supplementary Table 1 for the studies.

Variation in estimates of among-line variance within the same taxon and trait type: an experiment in *D. serrata*

To what extent do differences in magnitude among estimates of V_M reflect differences in mutation number and/or effect sizes (correlated with the above-investigated proxies of taxon and trait), vs factors such as mutations segregating within MA lines or environmental dependency of mutational effects? We conducted a further experiment to address this question. *Drosophila serrata* is a member of the *montium* species group, endemic to Australia and Papua New Guinea, which has been extensively used in quantitative genetic research, including study of mutational variance (e.g. [McGuigan and Blows 2013](#); [Hine et al. 2018](#); [Dugand et al. 2021](#)). A panel of 200 MA lines was founded from one of the *D. serrata* reference genome panel (DsRGP) lines described in [Reddiex et al. \(2018\)](#). These MA lines were each maintained by brother-sister inbreeding for 20 generations, following protocols established by [McGuigan et al. \(2011\)](#) to minimize selection. Genome-wide heterozygosity was very low (0.3%) in the DsRGP line that founded the MA lines, and among-line variance for wing traits (defined below) was not statistically supported in the first generation of the MA (S. Chenoweth, pers. comm.).

As detailed below, we applied an experimental design to this MA panel that allowed us to generate repeated estimates of the magnitude of among-line variance over 6 sequential generations, and characterize the relative contribution to differences among these sequential estimates of (1) mutations segregating within the MA lines or (2) unintentional variation in environment. We randomly chose 42 of the MA lines for this investigation based on the median number of MA lines in the reviewed published studies (see Results). The number of MA lines is the relevant degrees of freedom for the among-line variance, and this value (42) allows us to consider the other 2 effects against a relevant level of sampling error. Quantitative genetic parameters are associated with large sampling errors ([Klein et al. 1973](#); [Klein 1974](#); [Lynch and Walsh 1998](#)), and the relatively low signal (i.e. few genetic differences) among MA lines will make mutational variance particularly vulnerable to “noisy” estimation, and as such, it is important to document the potential for statistical sampling error to contribute to the observed variation among published estimates.

There were 3 key aspects of the experimental design that allowed us to test whether segregating variation or unintended environmental variation could explain differences among repeated estimates of among-line variance. First, we increased the population size within each MA line to a minimum of 12 males and 12 females ([Fig. 2a](#)). Empirical evidence suggests that population sizes as low as 10 may be sufficient to prevent fixation of mutations ([Estes et al. 2004](#); [Katju et al. 2015](#); [Luijckx et al. 2018](#)). Therefore, we expect no ongoing fixation of mutations among lines during this experiment, and for the repeated estimates of among-line variance to be true replicate sampling of the same mutations (but also test this assumption, as detailed below). We note that these changes in census population size complicate calculation of a per-generation rate of increase in phenotypic variance ([Lynch and Hill 1986](#); [Lynch and Walsh 1998](#)); here, we instead focus on the among-line variance, V_L , and do not interpret a per-generation rate of change.

The second key aspect of the experimental design was to manipulate the mutation-selection-drift dynamics within an MA line; this was achieved by imposing 2, substantially different, population sizes on sublines of each of the 42 MA lines ($N = 24$ vs

288 flies, referred to hereafter as small, S, and large, L, population size treatments: [Fig. 2a](#)). Segregating variants within MA lines (i.e. mutations that have not yet been fixed or lost) could cause transient inflation of among and/or within line variance (V_E), impacting on both the estimation and scaling of V_L , and this manipulation allowed us to determine the magnitude of this effect. The treatments contrast deterministic evolution of mutations with relatively strong ($s > \sim 0.038$: $N_e \sim 13$), vs weak ($s > 0.003$: $N_e \sim 158$) fitness effects, based on $s = 1/2N_e$ ([Wright 1931](#); [Kimura 1983](#)) and genomic estimates of N_e in MA lines of *D. melanogaster* maintained similarly to our small population size treatment (10 males and 10 females: [Huang et al. 2016](#)). The S and L treatments therefore had different opportunities for new mutations to increase in frequency within a line, and thus for the magnitude of within-line variance.

The final key aspect of the experimental design was the repeated measures themselves, allowing us to observe the effect of environmental variation on among-line variance. If the phenotypic effects of a mutation are context-dependent (i.e. exhibit G×E variance), then unintended differences in assay conditions could contribute heterogeneity among estimates when phenotypic data is collected at different timepoints (or in different laboratories). We randomly sampled the average environmental conditions present within our laboratory by repeatedly sampling the lines (genotypes) over 6 consecutive generations. Thus, our experiment consisted of applying 2 population size treatments (S, L) to each of 42 lines (derived from a classical MA experiment, with low among-line variation), where these 84 lines were maintained under the same conditions (12 flies per sex per vial founding each generation, with S and L differing in the number of vials) for 6 generations ([Fig. 2a](#)). As detailed below, we consider 11 wing shape and size traits. This allows us to understand the general influences of segregating variation, environment and sampling error for a set of related morphological traits. After accounting for the 3 factors that are the main focus of the investigation, we also determine whether the magnitude of V_L varies among these traits, allowing insight into potential magnitude of differences in mutational variance among traits within the same category (morphology).

Data collection

Each generation, 12 males (6 from each of 2 rearing vials) from each of the 84 sublines were randomly sampled for wing phenotypes ([Fig. 2](#)). Wings were mounted on microscope slides and photographed using a Leica MZ6 microscope camera and the software LAS EZ v2.0.0 (Leica Microsystems Ltd, Switzerland). A total of 5,135 wings were landmarked for 9 positions, defined by wing vein and margin intersections ([Fig. 2b](#)), using the software tpsDIG2 ([Rohlf 2013](#)). The number of wings were evenly distributed across treatments (2,583 in S and 2,552 in L) and generations (~425 per generation, per treatment). Landmarks were aligned using a General Procrustes fit in tpsRelw ([Rohlf 2007](#)). Centroid size (CS), the square root of the sum of squared deviations of the coordinates from the centroid ([Rohlf 1999](#)), was recorded as a metric of wing size. The aligned X-Y coordinates for each landmark were then used to calculate 10 inter-landmark distances (ILDs) ([Fig. 2b](#)). ILDs scores were re-scaled prior to analysis (multiplied by 100) to aid model convergence. Outliers >3.0 SD from the mean were removed for each of the 11 traits (10 ILD and size) (329 measures across the 56,485 total measures).

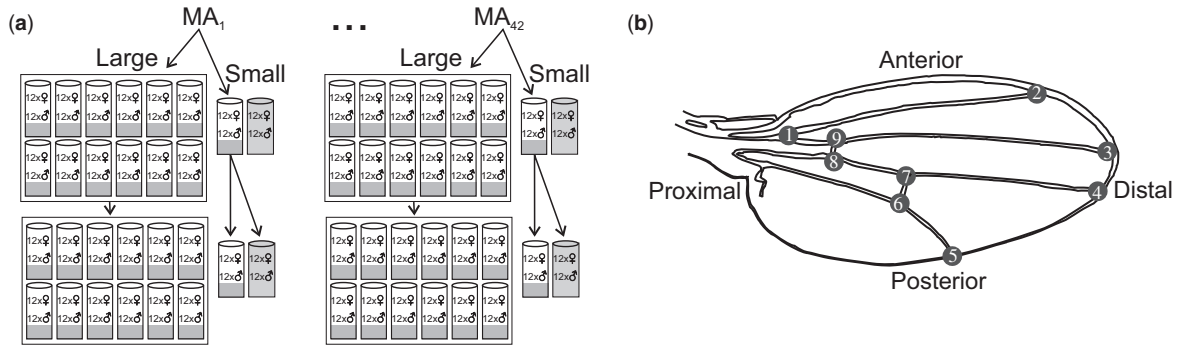


Fig. 2. Schematic of design (a) and phenotypes (b) from a manipulative experiment in *D. serrata*. (A) 42 MA lines (evolved through 20 generations of brother-sister mating) each founded 2 sublines: Small (S; 12 virgin males and 12 virgin females) and Large (L; 144 virgin males and 144 virgin females, distributed evenly among 12 vials). These 84 lines (S and L subline per 42 MA lines) were maintained at these census population sizes for 6 generations (only 2 shown here). Each generation, all emergent flies from the 12 vials per L subline were pooled prior to virgin collection. For S sublines, 2 vials were established each generation; the focal vial contributed offspring to the next generation, while the replicate vial (gray shaded) did not. Each generation, 1 wing was sampled from each of 6 randomly chosen males from each of 2 vials per line (focal and replicate vials for S; randomly chosen 2 for L). (B) Wing size and shape were characterized from landmarks recorded on an image of each wing: proximal (1) and distal intersections of the radial vein (2); distal intersections of medial (3), cubital (4), and distal (5) veins and; the posterior (6, 7) and anterior (8, 9) cross-veins. ILD traits were described by their endpoint landmarks (e.g. ILD1.2 was the distance between landmark 1 and landmark 2).

Analyses of variation in among-line variance estimates

Our experimental design allows us to repeatedly estimate variance among MA lines under conditions where we expect the number of mutations fixed among the lines, and their phenotypic effects, to be constant, and thus to investigate other potential causes of variability in estimates. We first treat the data from each generation and population size treatment as independent experiments of similar size (number of lines and individuals sampled per line) to typical MA experiments. To estimate among line variance from these 12 “experiments” for each of the 11 traits we fit the following model using REML in PROC MIXED in SAS v9.4 (SAS Institute Inc., Cary, NC.):

$$y_{klm} = \mu + \text{Line}_k + \text{Vial}_{l(k)} + \varepsilon_{klm} \quad (2)$$

where y_{klm} was the trait value for the m th wing (individual), from the l th vial, within the k th line, μ was the mean value of these observations; Line and replicate rearing Vial (nested within line) were fit as random effects, along with the among-individual variation (residual error, ε). We used REML-MVN sampling (Meyer and Houle 2013; Houle and Meyer 2015; Sztepanacz and Blows 2017) to estimate CIs, sampling 10,000 times from $N \sim (\hat{\theta}, \mathbf{V})$ using the morm in R [v. 3.6.1], where $\hat{\theta}$ was the vector of REML random effect parameter estimates, and \mathbf{V} was their inverse Fisher information matrix, $\mathbf{I}(\hat{\theta})^{-1}$. We similarly estimated the CIs for the trait mean, sampling based on least-squares mean and SE estimates output from model (2). The samples of random effect variances were not constrained to the parameter space (i.e. could be negative), allowing inference of statistical support when the lower 5% CI did not encompass zero (a 1-tailed test); this approach is equivalent to a log likelihood ratio test (LRT) (Dugand et al. 2021). Here, we are interested in general patterns of variability among these 12 “experiments,” and thus do not correct for multiple testing.

As detailed in the Results, substantial heterogeneity in magnitude was observed among the 12 replicate estimates of V_L per trait. We considered the potential contribution to this heterogeneity from unintentional heterogeneity in the culture conditions among sampling time points (generations) or between replicate measures of the same MA line within a generation (the S and L treatments). First, to determine if simple effects of variability in culture

conditions on trait scale could account for the variability of among-line (mutational) variation, we placed estimates on a heritability (V_L/V_E where V_E was the sum of among and within vial variances) or coefficient of variance ($100 \times \sqrt{V_L/\bar{X}}$) scales, and calculated confidence intervals by applying these equations to each of the 10,000 samples described above (and applying the sign correction for coefficients of variance as detailed in the meta-analysis methods). We further explored the relationship between V_L and the scaling parameters by regressing the 12 estimates of V_L on the corresponding estimates of V_E or trait mean.

Second, we determined whether mutational effects changed in response to the unintended changes in culture conditions, with such G×E causing differences among sequential estimates of V_L . Therefore, we extended this investigation, following García-Dorado et al. (2000) in treating different generations as different environments to formally test the null hypothesis that there was no G×E variance within the S or L treatments, using PROC MIXED and REML to fit:

$$y_{klm} = \mu + G_j + \text{Line}_k + (G \times \text{Line})_{jk} + \text{Vial}_{l(jk)} + \varepsilon_{klm} \quad (3)$$

where the fixed effect of generation (G) accounted for differences in trait mean among generations and the random effect of G(eneration) × Line estimated the variation in genetic effects among generations (where generations represent different local environments). For the component of V_E (i.e. Vial and residual), generation-specific effects were modeled (using the GROUP statement) to account for among-generation heterogeneity in the magnitude of V_E . This model was applied to each trait within each population size treatment, and statistical support for G × Line (and for Line) was determined using log-LRTs (0.5 d.f.: Self and Liang 1987; Littell et al. 2006) to compare model (3) to reduced models that did not fit G × Line (or did not fit Line). We applied the Benjamini-Hochberg method (Benjamini and Hochberg 1995) to correct for multiple hypothesis testing (within each random effect), using a conservative 5% false discovery rate (FDR). Sampling based on the REML variance estimate and the Fisher information matrix, as detailed above, was used to estimate CIs for plotting.

While nonzero generation by line variance could reveal the presence of environment-specific mutational effects, it could alternatively be explained by changes in the frequency at which mutations

were segregating within or among lines. In contrast to environmental heterogeneity, we expect these evolutionary processes to systematically differ between the 2 population size treatments due to the different efficacy of selection in the S vs L sublines, and the independent sampling of mutations in the sublines after they were established. Differences between S and L are predicted to increase with increasing time since divergence. For each of the 11 traits, we analyzed all data (from both L and S) collected within a single generation, using PROC MIXED and REML to fit:

$$y_{ijklm} = \mu + \text{Treat}_i + \text{Line}_k + \text{Vial}_{l(ijk)} + \varepsilon_{ijklm}, \quad (4)$$

where treatment was fit as a fixed effect to account for differences in trait mean between L and S panels of sublines within that generation. Vial and residual are as described for model (2). At the among-line level, we took advantage of the paired subplot design to model the between treatment variance-covariance matrix. We employed LRT to test 2 hypotheses. First, we determined whether, for these analyses within a generation, there was support for differences between treatments in the magnitude of V_E . Mutations that are segregating (i.e. occur at frequencies other than 0 or 1) within an MA line will contribute to variation both among-vials and the residual. We compared a model in which one (common to both Treatments) among Vial variance and one residual variance were estimated to a model in which Treatment-specific variances were modeled at both levels (fit using the GROUP statement). Second, we tested whether the 2 copies of the MA lines had diverged from one another by testing the hypothesis that the correlation between the paired sublines was <1.00 (implemented using a PARMS statement). To correct for multiple hypothesis testing (within each hypothesis), we employed a FDR correction as described above.

Finally, as there was little support for varying mutational effects (no $G \times E$) or number (no divergence between S and L) contributing to the apparent heterogeneity among repeated estimates per trait (detailed in Results), we use our data to revisit the question of whether traits inherently differ from one another in the magnitude of mutational variance. We obtained a single estimate of V_L per trait by using PROC MIXED in SAS to fit:

$$y_{ijklm} = \mu + T_i + G_j + TG_{ij} + \text{Line}_k + (G \times \text{Line})_{jk} + \text{Vial}_{l(ijk)} + \varepsilon_{ijklm}, \quad (5)$$

where all effects are as described above, including the fixed effects of population size treatment (T), generation (G), and their interaction (TG), as well as the random effects of Line, Generation by Line, Vial and residual. We obtained REML-MVN CIs for each parameter, as described above. To test whether observed differences in V_L among traits were due to differences among them in scale, we took the among-line (V_L) estimates from model (5) and regressed them on the corresponding estimates of environmental variance or on the squared trait mean. These regressions were applied to the REML parameter estimates, and to each of 10,000 samples of these parameters to determine statistical significance (95% CI of slope did not include zero).

Results

Meta-analysis of published mutational variance estimates

Our final meta-analysis data set consisted of 154 estimates of h_M^2 and 148 estimates of CV_M . These estimates of h_M^2 ranged from

2.50×10^{-5} to 1.02×10^{-2} , while CV_M ranged from 0.13 to 7.32. We predicted that differences in genome size and/or genomic mutation rate would cause differences in the magnitude of mutational variance among taxa. However, there was no statistical support for a difference in mutational variance among the taxon categories (h_M^2 : $F_{3,24.2} = 2.28$, $P = 0.1044$; CV_M : $F_{3,17.5} = 1.24$, $P = 0.3261$), although h_M^2 estimates from Plants were markedly lower than estimates from *Daphnia* and *Drosophila* (Fig. 3a). We predicted that differences among traits in the number of contributing loci would cause differences in the magnitude of mutational variance. However, trait categories only differed in the magnitude of CV_M ($F_{5,37.1} = 3.86$, $P = 0.0064$), not h_M^2 ($F_{5,85.9} = 0.40$, $P = 0.8497$) (Fig. 4). Overall, these factors (taxon and trait category) accounted for 1.64% of the variation in estimates of h_M^2 and 9.88% of variation among CV_M estimates.

Although not statistically supported, it is notable that the among-trait trend did not follow predictions for h_M^2 : fitness traits had the largest average h_M^2 , not the lowest as expected (Fig. 4a). Following Houle et al. (1996), we also analyzed V_E (fit model (1) to $CV_E = \frac{\sqrt{V_E}}{\bar{x}}$). There was no statistical support for a difference among traits in the magnitude of CV_E ($F_{5,25.2} = 1.32$, $P = 0.2887$); morphology (average $CV_E = 7.4$) and physiology (71.6) differed the most, with life history traits having intermediate values (e.g. fitness = 42.6) (Supplementary Fig. 2a). For CV_M , the statistically supported differences did follow the predicted pattern, with morphological traits having the smallest CV_M and fitness the largest (Fig. 4b). Survival notably had lower CV_M than productivity and fitness (Fig. 4b), although surviving to reproduce was a component of fitness. Physiological traits had a similar magnitude of CV_M to morphological traits, lower than any life history trait category (Fig. 4b).

While the lack of observed difference in scaled estimates of V_M among taxa may reflect a true commonality among species in this important evolutionary parameter, aspects of the MA design also differed markedly among taxa. Estimates from Plants were derived from MA experiments that were of short duration (maximum 25 generations) relative to other taxa (median 44, 75, and 214 for *Drosophila*, *Daphnia* and *Nematode*, respectively) (Supplementary Fig. 2b). As mutations arise independently in each MA line, the number of MA lines maintained may also influence the number of mutations that similar duration MA could sample, and the potential for sampling of rarer mutational effects (i.e. from the tails of the distribution of effect sizes) to influence estimates; while the median number of MA lines was similar in *Nematode* (43) *Plant* (50) and *Drosophila* (52), it was substantially lower in *Daphnia* (8) (Supplementary Fig. 2c). While mutational variance was estimated for multiple types of traits in every taxon, most data from Plants was for fitness, while most data from *Daphnia* was for morphological traits, and *Drosophila* and *Nematode* were the only 2 taxon categories that contributed estimates for physiological traits (Supplementary Fig. 2d).

Variation in estimates of among-line variance within the same taxon and trait type

Within the *D. serrata* experiment, we first determined the heterogeneity in among-line variance (V_L) under the assumption that mutation number and effects were constant, treating the 12 V_L estimates per trait as independent MA experiments. There was substantial variation among the 12 estimates per trait, with some differences of over an order of magnitude (Fig. 5; Supplementary Table 4). Notably, the smallest estimate of V_L for size (CS) was four times lower than the largest estimate, comparable to the 4-times difference among reported estimates of h_M^2 for body size in

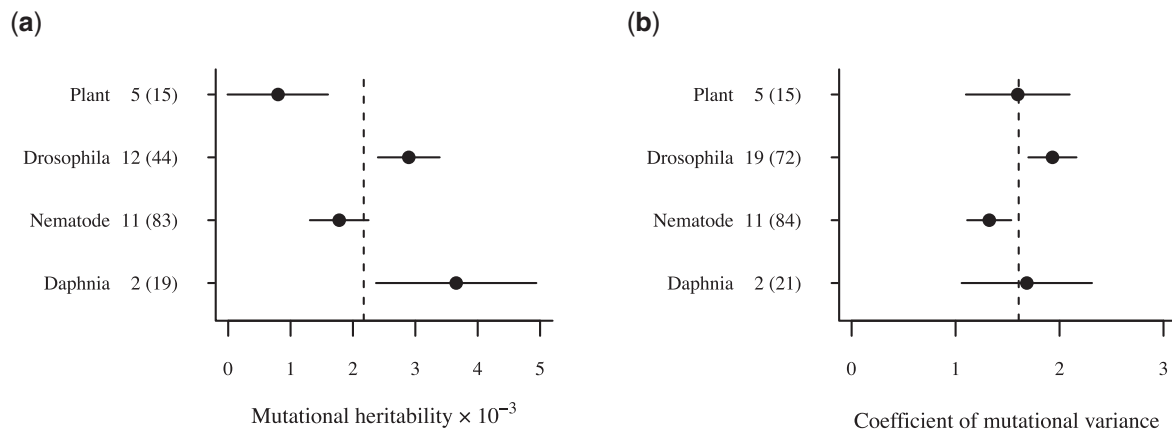


Fig. 3. Variation in estimates of (a) mutational heritability and (b) coefficient of mutational variance across taxon categories. Plotted are the least-squares mean estimate (\pm SE) from the analyses of model (1). The number of studies (and estimates) analyzed for each category are shown. The dashed line indicates the global mean value.

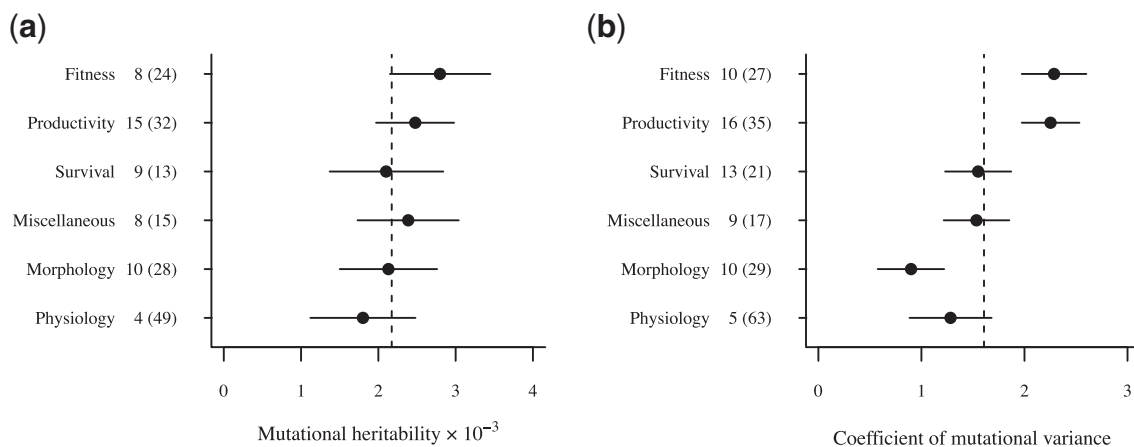


Fig. 4. Variation in estimates of (a) mutational heritability and (b) coefficient of mutational variance across trait categories. Plotted are the least-squares mean estimate (\pm SE) from the analyses of model (1). The number of studies (and estimates) analyzed for each category are shown. The dashed line indicates the global mean value.

C. elegans (Azevedo et al. 2002; Estes et al. 2005; Ostrow et al. 2007). Predictably given this heterogeneity in magnitude of effect (i.e. in V_L), there was also inconsistent statistical support for the presence of V_L for most traits, despite consistent sample sizes in each of the 12 “experiments” (Fig. 5; Supplementary Table 4). Thus, we might draw very different conclusions about the magnitude of V_L for a trait, depending on which “experiment” we had conducted (Fig. 5).

Due to the changes in N_e within this experiment, we do not place these V_L estimates on a per-generation scale (i.e. do not calculate V_M). However, there is no trend for V_L to increase through time (i.e. no signal of ongoing divergence through fixation of mutations), or to diverge between the different population size (N_e) treatments (Fig. 5) (addressed further below). Therefore, calculating V_M is not expected to eliminate the heterogeneity in estimates.

Reporting mutational variance estimates as h_M^2 or CV_M facilitates comparison among estimates by accounting for inherent differences in scale. Although here the 12 estimates come from the same trait, scale differences may still arise through typical effects of any unintended variation in culture conditions (occurring among generations or between the replicate S and L sub-lines) on nongenetic trait variance (V_E) or mean. Both V_E and the trait mean varied substantially among the 12 repeated estimates

for all traits (Supplementary Figs. 3 and 4; Supplementary Table 4). However, this variation in V_E and trait mean was independent of the observed variation in V_L ; regressing the 12 estimates of V_L on their corresponding estimate of V_E or trait mean supported only 1 slope (ILD3.7, V_L on mean) as statistically different from zero (although this did not remain significant following FDR correction) (Fig. 6; Supplementary Table 5). Consistent with this pervasive independence of V_L from the scaling factors for these repeated measures of the same trait, when the 12 estimates were placed on either a heritability (Supplementary Fig. 5; Supplementary Table 4) or coefficient of variance scale (Supplementary Fig. 6; Supplementary Table 4), the variation among them was of a similar magnitude to that observed for V_L itself (i.e. the variation plotted in Fig. 5). Plotting the scaled estimates (h^2 or CV) against their respective numerator (V_L or $\sqrt{V_L}$) and denominator (V_E or mean) illustrates the predominant contribution from variation in V_L to variation in the scaled estimates (Supplementary Fig. 7). Overall, the 12 estimates of V_L are more variable than the corresponding estimates of V_E or trait mean, with variability of the scaled estimates (h^2 or CV) more similar to V_L than to their respective scaling factor (Supplementary Fig. 8).

To compare the variability among published estimates of h_M^2 and CV_M in similar morphological traits (excluding bristle traits: 19 estimates, 8 each from *Daphnia* and *Nematode*, 3 *Drosophila*

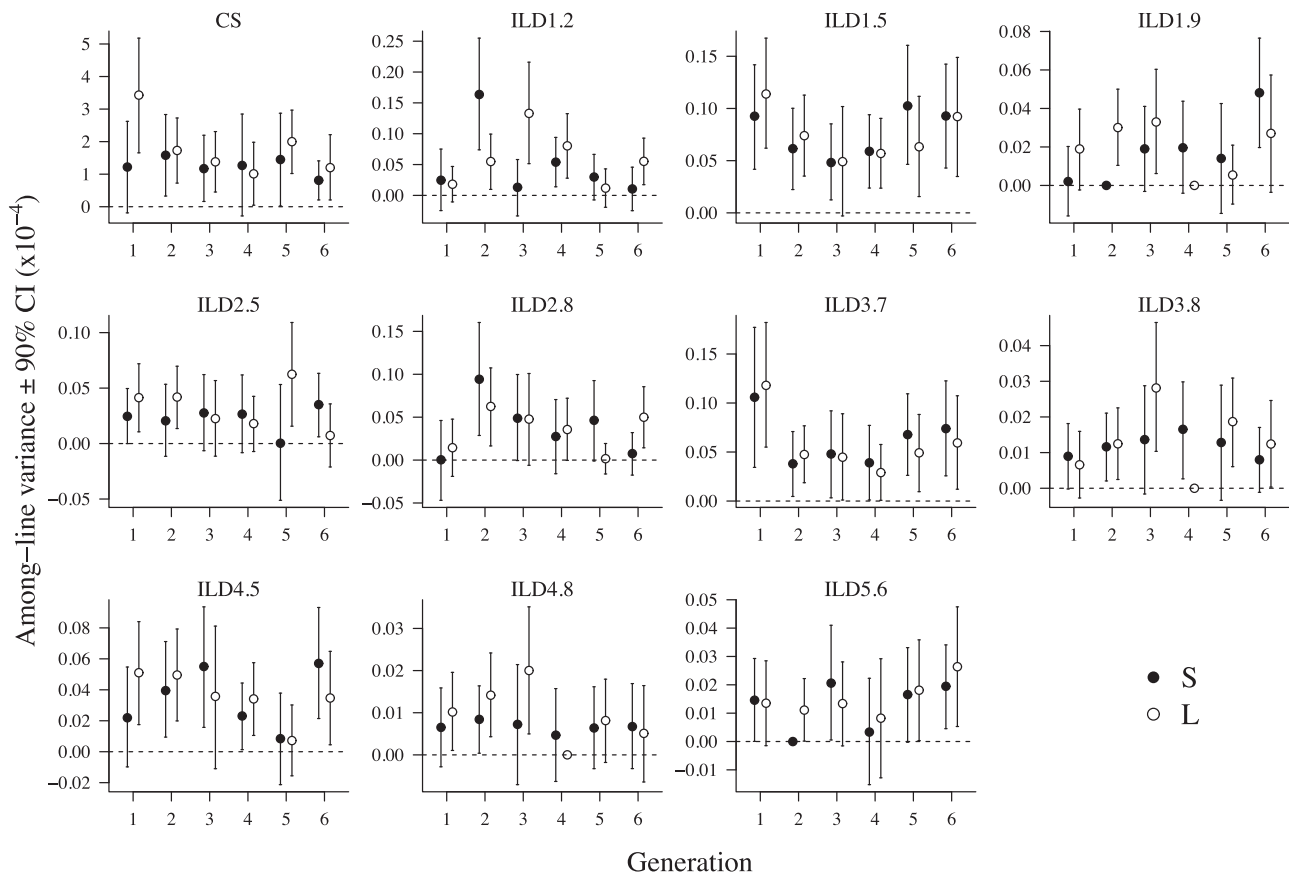


Fig. 5. Among-line variance estimates across 6 generations in an experiment in *D. serrata*. Variances were estimated independently for each trait (panel; see Fig. 2b for trait definitions) in each generation (x-axis) for each of the 2 population size treatments (Small: solid circles; Large: open circles). Plotted are the REML point estimate, and the REML-MVN 90% CIs. The dashed horizontal line indicates 0; estimates for which the lower CI did not overlap zero were interpreted as statistically supported. Where REML estimates of among-line variance were zero, CI were not estimated.

and 1 Plant) to the variability among the V_L estimates for *D. serrata* wing trait traits, we calculated the coefficient of variance ($cv = \text{standard deviation}/\text{mean}$) for each set of estimates. The cv of all 19 published h_M^2 (0.80) and CV_M (0.82) estimates was above the median cv of the 11 *D. serrata* traits on the observed V_L scale (0.55), V_E -scale (h^2 : 0.54) or mean-scale (CV : 0.39), but nonetheless within the same range: cv of V_L (V_E -scale; mean-scaled) ranged from 0.30 (0.37; 0.15) up to 0.92 (0.85; 0.62) across the 11 traits (Supplementary Fig. 8). Within taxa, cv of published h_M^2 (CV_M) estimates ranged from 0.29 (0.08) for the 3 *Drosophila* estimates up to 0.97 (0.38) in *Daphnia*, with a median of the 3 within-taxon (*Drosophila*, *Daphnia*, and *Nematode*) cv s of 0.64 (0.24). Thus, overall, the heterogeneity among repeated V_L estimates in *D. serrata* is of a similar magnitude to the variation among published estimates of the same trait type.

Having established that variation in the magnitude of V_L is not a simple consequence of varying scale (V_E or mean), we investigated other putative causes. In addition to the general effects on scale, unintended differences in culture conditions among generations could also affect V_L if mutations had context-dependent effects on the trait (i.e. $G \times E$), as characterized by generation by among-line variance. $G \times E$ was statistically supported in only 4 cases, with only 2 remaining significant at a 5% FDR (LRT for CS, in L: $\chi_1^2 = 13.9$, $P = 0.0001$; LRT for ILD2.5 in L: $\chi_1^2 = 11.1$, $P = 0.0005$) (Fig. 7a). Thus, for these 2 traits, the analysis suggests that mutational effects, and the magnitude of among-line variance, may depend on the specific conditions under which the traits were assayed.

Ongoing mutation-drift-selection processes could contribute to variation among the 12 estimates, where the S and L treatments are expected differ in the potential effects of these processes on both within and among-line variance. Segregating variation within a line will contribute to the estimate of V_E , and we determined whether the S and L treatments differed in the magnitude of V_E , analyzing each of the 11 traits within each of the 6 generations separately. Eight of the 66 estimates of V_E differed significantly between S and L at $P < 0.05$, but only one remained significant at 5% FDR (CS in generation 5; Supplementary Table 6). There was no statistical support for the S and L sublines founded by each of the 42 original MA to have diverged from one another in the mutations they carried, either through initial sampling when lines were founded, or through (near) fixation of mutations arising after establishment of the sublines. Specifically, the among-line correlation between S and L sublines was not statistically distinguishable from 1.0 for any trait in any generation (Supplementary Table 6).

Finally, we obtained a single estimate of among-line variance for each trait to determine whether the magnitude of V_L was consistent among the 10 wing shape traits, which are expected to share a genetic basis, and developmental pathways (e.g. Mezey et al. 2005; Neto-Silva et al. 2009). The among-trait heterogeneity (i.e. nonoverlapping CIs: Fig. 8a; $cv = 0.77$) was larger than the median variability among the repeated estimates per trait ($cv = 0.55$; see above), and comparable to the variability among published estimates of mutational variance in morphological traits ($cv = 0.80$, detailed above). Variation among shape traits

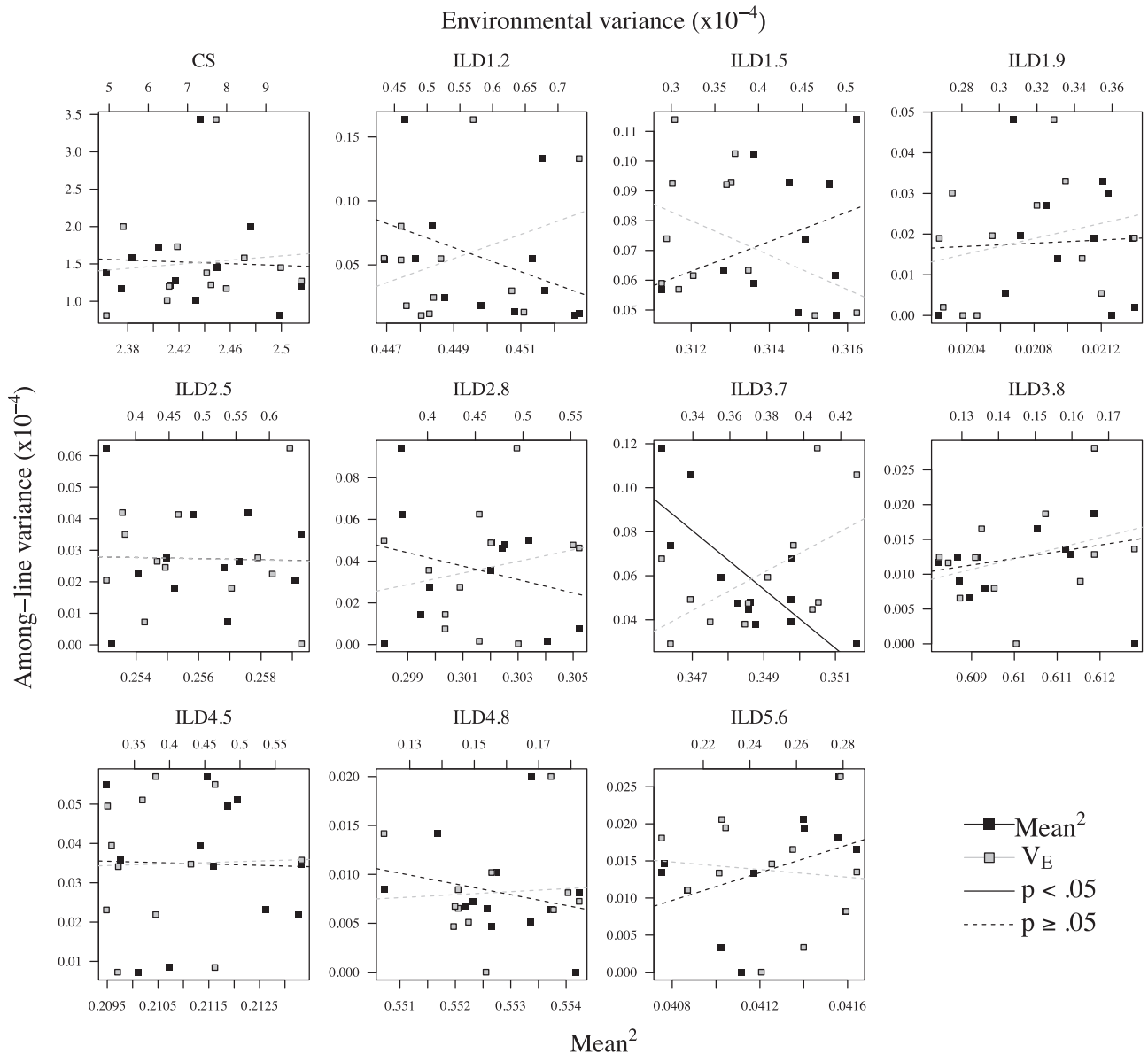


Fig. 6. Among-line variance estimates for *D. serrata* wing traits plotted as a function of trait mean or variance. The 12 estimates of among-line variances for each of the 11 wing traits (panels; see Fig. 2b for trait definitions) are plotted against the corresponding (i.e. same generation and treatment) squared trait mean (bottom x-axis, black symbols) or environmental variance (summed among and within vial variances; top x-axis, gray symbols). All regression statistics are reported in Supplementary Table 4; only the effect of Mean^2 on V_L of ILD3.7 was significant at $P < 0.05$, although it does not remain significant after applying a 5% FDR correction.

(excluding size) in V_E accounted for $\sim 18\%$ (95% CI: 0.049–0.383) of this variation ($\beta = 0.064$ [95% CI: 0.050–0.106]), but variation in trait mean did not account for any ($\beta = 5.3 \times 10^{-7}$ [-0.74×10^{-8} – 1.86×10^{-6}]; $R^2 = 0.003$ [95% CI: <0.001 –0.027]) (Fig. 8b). Establishing whether these differences are informative of the inherent genetic architecture, or are a manifestation of the stochastic nature of mutation, requires repeating the estimation using either the same or a different genetic background to determine if consistent differences among traits persist. When wing size was also considered, overall scale influenced V_L , with much of the variation among estimates accounted for by the scaling factors (V_E : $R^2 = 0.995$ [0.978–0.998]; Mean^2 : $R^2 = 0.921$ [0.903–0.925]), suggesting that there was little difference in the magnitude of underlying mutational variance between wing size and the shape traits (Fig. 8b).

Discussion

Although numerous estimates of mutational variance have been published, it remains unclear what contributes to the ~ 2 orders of magnitude difference among these estimates. Our meta-analytic investigation provided some support for a difference among trait types in the magnitude of mutational variance, but also revealed substantial confounding between potential causal factors. Analyses of data from a manipulative experiment in *D. serrata* suggests that, for the morphological traits under consideration, factors such as unintended heterogeneity in environmental conditions or transient segregation of mutations within MA lines may contribute little to the variation among estimates. Given this experimental design, and the evidence that mutation number and effect did not typically cause differences among repeated estimates, we conclude that substantial

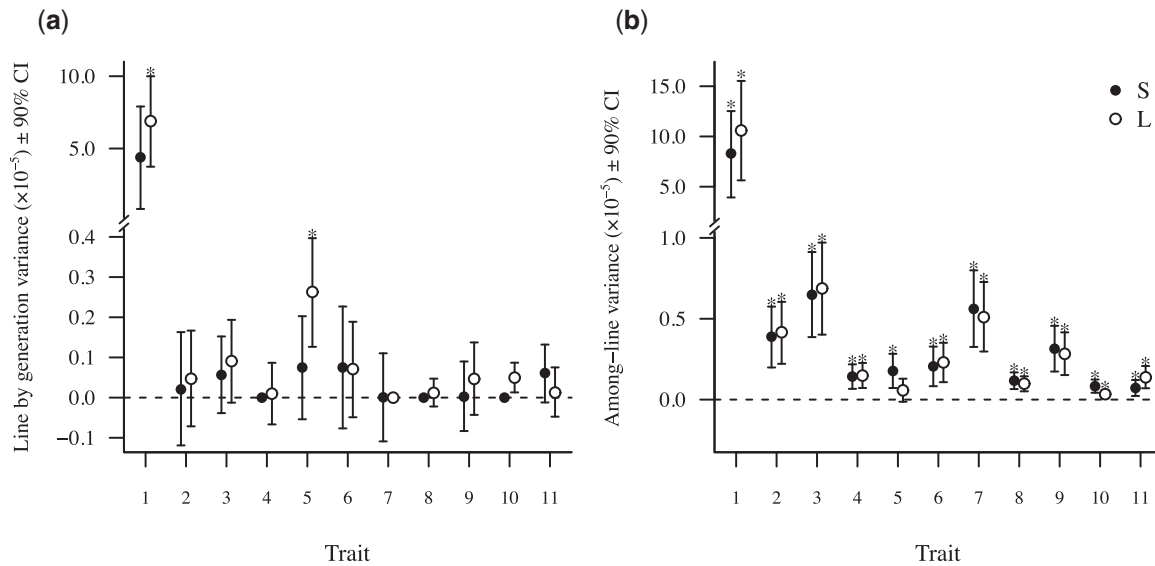


Fig. 7. Estimates of variance from an experiment in *D. serrata*. a) Among-line by generation ($G \times E$) variance and (b) among-line variance estimated for 11 *D. serrata* wing traits (x-axis), in 2 different population size treatments (Small: solid circles; Large: open circles). Plotted are the REML point estimates (from model 3) and the REML-MVN 90% CIs. The dashed horizontal line indicates 0; statistical significance was inferred where the lower 5% CI did not overlap 0. After applying a conservative 5% FDR correction, 2 estimates in a) and 21 in b) remained significant (indicated by an asterisk).

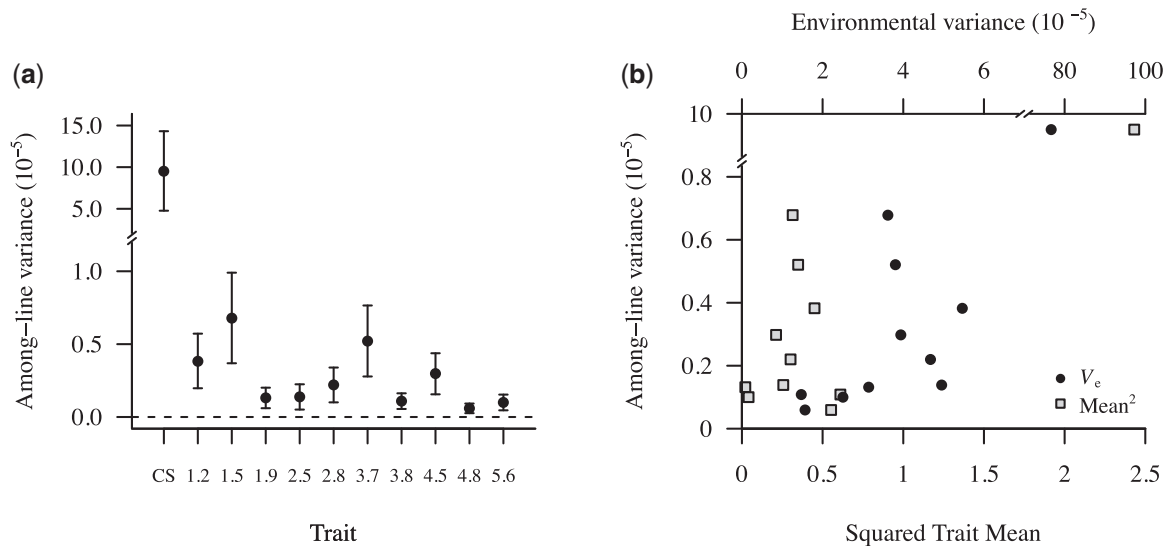


Fig. 8. Among-line variances for 11 wing traits in *D. serrata*. a) Among-line variance, V_L REML estimates (and 90% CI) (model 5; see Fig. 2b for trait definitions) are plotted. Dashed line indicates 0. b) REML estimates of among line variance (points in panel a) were plotted against the corresponding estimate of environmental variance (black circles, top x-axis) or mean squared (gray squares, bottom x-axis) of the trait. Regression results are reported in text.

variability among repeated estimates of the among-line variance must reflect sampling error. Below we discuss the specific outcomes and limitations of both approaches, and the implications our analyses have for future work characterizing mutational input to quantitative genetic variation.

Effects of taxon and trait type on the magnitude of mutational variance

Given the ~4-times higher per site mutation rate (Katju and Bergthorsson 2019), and slightly larger genome of *A. thaliana* relative to *C. elegans* we predicted (assuming the same mutational effect sizes) ~5 times more mutational variance in the Plant than Nematode taxon categories. However, the meta-analysis did not support a difference among taxa in the magnitude of mutational

variance, and the observed (strong but nonsignificant) pattern in h_M^2 contradicted this rank prediction, with Plants (to which *Arabidopsis* contributed most estimates) having substantially smaller h_M^2 than other taxa (Fig. 3a). Taxon categories differed substantially in the number of generations (Supplementary Fig. 2b), and the number of genomes (MA lines) (Supplementary Fig. 2c) sampled. However, scaling predicted genomic mutation by this opportunity for mutation also fails to predict the trend, with *Daphnia* MA experiments predicted to sample the fewest mutations but observed to have the largest h_M^2 (Fig. 3b). We suggest that further MA experiments, decoupling the confounded effects of MA duration and trait type from taxon, are warranted to determine whether V_M does vary among taxa. Advances in accessibility of genome data provides substantial scope for such

experiments to explicitly estimate relevant genomic parameters (e.g. frequency spectra for different types of mutations across putatively causal genes) alongside the phenotypic variation generated by those mutations (Katju and Bergthorsson 2019). Furthermore, given evidence that epigenetic mutations arise more frequently than genetic mutations (e.g. van der Graaf et al. 2015; Beltran et al. 2020), we suggest that the potential contribution of epimutations to patterns of heterogeneity of V_M should be explicitly assessed in future studies.

Houle et al. (1996; see also Lynch and Walsh 1998; Lynch et al. 1999) concluded that life-history traits had lower h_M^2 and higher CV_M than morphological traits, a pattern that is also observed in standing genetic variation (e.g. Houle 1992; Hansen et al. 2011; but see Hoffmann et al. 2016). However, this conclusion was not supported by our analysis, where the trend was for fitness and productivity to have the highest h_M^2 (nonsignificant) as well as highest CV_M (Fig. 4). As expected given this shared pattern, h_M^2 and CV_M estimates were positively correlated (Spearman's correlation coefficient: 0.309, $N=117$, $P=0.0007$). Although with reverse rank (life-history traits having lowest values), standing genetic variation estimates have also been reported to be positively correlated between the 2 scales (h^2 and CV) when biologically uninformative CV estimates were excluded (Hoffmann et al. 2016). Garcia-Gonzalez et al. (2012) highlight the potential for skewed data distributions to inflate (deflate) CV , an issue that may be particularly relevant to estimates of CV_M . While strong bias toward mutations that decrease mean fitness has been reported (Halligan and Keightley 2009), bias in other traits is less well-established. If trait types differ in the magnitude of directional bias of mutational effects, this may also result in differences in skew, and exaggerate differences between trait types on the CV_M scale. Again, resolution of the key question of whether differences among traits in h_M^2 and CV_M reflect differences in mutation number and/or effect size may depend on further genomic data.

The contributions of unintended environmental variation, mutation-drift-selection processes, and sampling error to variation in the magnitude of mutational variance

We observed substantial variation among repeated estimates of V_L each of 11 wing traits measured in *D. serrata* (Fig. 5), resulting in variation among scaled (h^2 or CV) estimates that was of comparable magnitude to the differences observed among published estimates. Although both V_E and trait mean also varied among repeated measures (Supplementary Figs. 3 and 4), this heterogeneity was substantially less than that observed for V_L (or h^2 or CV) (Supplementary Fig. 8). Given the evidence that mutational effects can vary among environments (Kondrashov and Houle 1994; Martin and Lenormand 2006), we tested the effects on the magnitude of V_L resulting from unintentional and undocumented minor changes in culture conditions (e.g. density, humidity, or temperature), such as may occur among phenotype assays conducted at different times or in different laboratories. Variation in mutational effects among phenotypic assays (generations) was supported in only 2 cases (Fig. 7a). Garcia-Dorado et al. (2000) also found evidence of Gx E among consecutive generations for 1 (sternopleural bristle count) of 4 traits investigated. Notably, in *D. serrata*, wing size, which might be particularly sensitive to variation in energy availability (or competing energetic demands) (Cavicchi et al. 1985; Bitner-Mathé and Klaczko 1999), exhibited the strongest Gx E (Fig. 7a). Our results, and those of Garcia-Dorado et al. (2000) suggest that changes in mutational effects with

environment may contribute to heterogeneity among published estimates of some traits, which may reflect differences in trait environmental sensitivity, or potentially in the covariation of environmental sensitivity and mutational effect size (Lynch et al. 1999; Garcia-Dorado et al. 2000).

The mutation-drift process itself may also contribute to variability among published estimates due to effects on both the within-line variance (transient inflation leading to increased magnitude of V_E but not V_L) and among-line variance (transient contribution to V_L of additive or dominant mutations that are subsequently lost via random sampling). We introduced an \sim order of magnitude difference in census size in paired sets of MA sublines to manipulate the mutation-drift-selection processes. However, analyses did not support an effect of population size on either within- or among-line variation; size was again an exception, where there was some evidence that relaxed selection allowed the S treatment to accumulate greater within-line variance (Supplementary Table 6). The effect of segregating variation can be expected to be greater at smaller population sizes (e.g. when $N=2$, mutations can reach within-line frequency of 75% before being lost by drift) than considered here, and so may play a greater role in explaining variation among estimates from classical MA breeding designs. But, nonetheless, this factor did not account for the substantial heterogeneity that was observed among the 12 estimates per trait within the current study.

Rejecting general contributions from environmental variation and transient segregation of mutations as explanations of the heterogeneity among the 12 repeated estimates of V_L for the wing shape traits, we conclude that the observed variability is largely the consequence of sampling error. Lynch et al. (1999) suggested that a substantial part of the order of magnitude range of h_M^2 reported for *D. melanogaster* may be due to sampling error. Here, we observed the magnitude of heterogeneity among the 12 repeated estimates of V_L to be similar to the heterogeneity among published, scaled estimates of mutational variance in morphological traits, consistent with their prediction. We observed that V_L estimates varied markedly more than the other estimated parameters (Supplementary Fig. 8), as expected given that quantitative genetic parameters are associated with relatively large sampling errors. Notably, V_E was more variable among the 12 estimates than trait mean was, which may lead to greater variability among estimates of h_M^2 than CV_M (Supplementary Fig. 8). We designed this experiment to mimic an average MA sample size, and considered traits expected to have relatively low experimental noise (residual variation) and small effect size (due to the relatively few generations; see e.g. Vassilieva et al. 2000). While traits and MA panels will vary in their vulnerability to sampling error, we nonetheless suggest that greater consideration must be given to the consequences of this error when designing experiments. The heterogeneity among repeated estimates resulted in the total confidence range for each trait spanning a far greater region than suggested by the error estimated for each repeated estimation of V_L (Fig. 5), indicating that within-study estimates of error do not fully capture the uncertainty in estimates.

The sequential repeated-measures experimental design provided greater statistical control over the experimental noise, allowing us to consistently detect statistically significant mutational variance in all traits (Fig. 8a), including in traits for which very few of the 12 estimates were distinguishable from 0 (e.g. ILD2.8; Fig. 5). While increasing sample sizes within a generation is likely to have similarly improved estimate precision, this can be logistically prohibitive in some systems. Given these limits, our analysis highlights the potential benefits of short-term

repeated measures (sequential generations) to improve estimate precision, and power to detect small effects. Repeated measures of lines at relatively large generation intervals have also been utilized to estimate V_M as the slope of the regression of among-line variance on generation (Vassilieva et al. 2000; Houle and Nuzhdin 2004; McGuigan et al. 2011), which may also improve estimation.

Understanding the contribution that mutations make to evolutionary and genetic phenomena relies on accurate estimates of the phenotypic variance generated by new mutation. Our meta-analysis of empirical estimates of mutational variance was unsuccessful in clearly resolving causes of variation due to confounding of predictors, and inconsistent patterns. Our manipulative experiment suggested that sampling error may contribute substantially to estimate variability, and demonstrated that repeated measures over few (e.g. sequential) generations provides a simple but effective approach to address this and improve inference. Overall, further empirical studies are needed to fully assess how both general and study specific factors influence V_M estimates, where improved precision and replicability in estimates will consequently advance broader evolutionary questions such as those addressing the maintenance of quantitative genetic variance (Barton and Turelli 1989; Johnson and Barton 2005; Walsh and Lynch 2018).

Data availability

Both analyzed datasets are available at doi: 10.6084/m9.figshare.14913051.

[Supplemental material](#) is available at GENETICS online.

Acknowledgments

The authors thank Stephen F. Chenoweth for providing us with the *D. serrata* MA lines, and Adam Reddiex, Nicholas Appleton, Jack Price and Derek Sun for their help with maintaining the flies and collecting the data. The authors also thank Jan Engelstädter for suggesting the combinational approach, and Emma Hine for contributions to preparing figures. This manuscript was improved by the input of 3 anonymous reviewers and the AE. This research was funded by the Australian Research Council.

Conflicts of interest

None declared.

Literature cited

- Azevedo RBR, Keightley PD, Laurén-Määttä C, Vassilieva LL, Lynch M, Leroi AM. Spontaneous mutational variation for body size in *Caenorhabditis elegans*. *Genetics*. 2002;162(2):755–765.
- Baer CF. Quantifying the decanalizing effects of spontaneous mutations in *Rhabditid nematodes*. *Am Nat*. 2008;172(2):272–281.
- Baer CF, Joyner-Matos J, Ostrow D, Grigaltchik V, Salomon MP, Upadhyay A. Rapid decline in fitness of mutation accumulation lines of gonochoristic (outcrossing) *Caenorhabditis nematodes*. *Evolution*. 2010;64(11):3242–3253.
- Barton NH, Keightley PD. Understanding quantitative genetic variation. *Nat Rev Genet*. 2002;3(1):11–21.
- Barton NH, Turelli M. Evolutionary quantitative genetics: how little do we know? *Annu Rev Genet*. 1989;23:337–370.
- Beltran T, Shahrezaei V, Katju V, Sarkies P. Epimutations driven by small RNAs arise frequently but most have limited duration in *Caenorhabditis elegans*. *Nat Ecol Evol*. 2020;4(11):1539–1548.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B Statist Methodol*. 1995;57(1):289–300.
- Besnard F, Picao-Osorio J, Dubois C, Felix MA. A broad mutational target explains a fast rate of phenotypic evolution. *Elife*. 2020;9:e54928.
- Bitner-Mathé BC, Klaczko LB. Plasticity of *Drosophila melanogaster* wing morphology: effects of sex, temperature and density. *Genetica*. 1999;105(2):203–210.
- Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. *Cell*. 2017;169(7):1177–1186.
- Braendle C, Baer CF, Félix M-A. Bias and evolution of the mutationally accessible phenotypic space in a developmental system. *PLoS Genet*. 2010;6(3):e1000877.
- Cavicchi S, Guerra D, Giorgi G, Pezzoli C. Temperature-related divergence in experimental populations of *Drosophila melanogaster*. I. Genetic and developmental basis of wing size and shape variation. *Genetics*. 1985;109(4):665–689.
- Dugand RJ, Aguirre JD, Hine E, Blows MW, McGuigan K. The contribution of mutation and selection to multivariate quantitative genetic variance in an outbred population of *Drosophila serrata*. *Proc Natl Acad Sci USA*. 2021;118(31):e2026217118.
- Estes S, Ajie BC, Lynch M, Phillips PC. Spontaneous mutational correlations for life-history, morphological and behavioral characters in *Caenorhabditis elegans*. *Genetics*. 2005;170(2):645–653.
- Estes S, Phillips PC, Denver DR, Thomas WK, Lynch M. Mutation accumulation in populations of varying size: the distribution of mutational effects for fitness correlates in *Caenorhabditis elegans*. *Genetics*. 2004;166(3):1269–1279.
- Fry JD, Heinsohn SL, Mackay TF. The contribution of new mutations to genotype-environment interaction for fitness in *Drosophila melanogaster*. *Evolution*. 1996;50(6):2316–2327.
- García-Dorado A, Fernández J, López-Fanjul C. Temporal uniformity of the spontaneous mutational variance of quantitative traits in *Drosophila melanogaster*. *Genet Res*. 2000;75(1):47–51.
- García-Gonzalez F, Simmons LW, Tomkins JL, Kotiaho JS, Evans JP. Comparing evolvabilities: common errors surrounding the calculation and use of coefficients of additive genetic variation. *Evolution*. 2012;66(8):2341–2349.
- Hall DW, Mahmoudizad R, Hurd AW, Joseph SB. Spontaneous mutations in diploid *Saccharomyces cerevisiae*: another thousand cell generations. *Genet Res (Camb)*. 2008;90(3):229–241.
- Halligan DL, Keightley PD. Spontaneous mutation accumulation studies in evolutionary genetics. *Annu Rev Ecol Evol Syst*. 2009;40(1):151–172.
- Hansen TF, Pelabon C, Houle D. Heritability is not evolvability. *Evol Biol*. 2011;38(3):258–277.
- Hine E, Runcie DE, McGuigan K, Blows MW. Uneven distribution of mutational variance across the transcriptome of *Drosophila serrata* revealed by high-dimensional analysis of gene expression. *Genetics*. 2018;209(4):1319–1328.
- Ho EKH, Macrae F, Latta LC, McIlroy P, Ebert D, Fields PD, Benner MJ, Schaack S. High and highly variable spontaneous mutation rates in *Daphnia*. *Mol Biol Evol*. 2020;37(11):3258–3266.
- Hoffmann AA, Merila J, Kristensen TN. Heritability and evolvability of fitness and nonfitness traits: lessons from livestock. *Evolution*. 2016;70(8):1770–1779.
- Houle D. Genetic covariance of fitness correlates: what genetic correlations are made of and why it matters. *Evolution*. 1991;45(3):630–648.

- Houle D. Comparing evolvability and variability of quantitative traits. *Genetics*. 1992;130(1):195–204.
- Houle D. How should we explain variation in the genetic variance of traits? *Genetica*. 1998;102:241–253.
- Houle D, Bolstad GH, van der Linde K, Hansen TF. Mutation predicts 40 million years of fly wing evolution. *Nature*. 2017;548(7668):447–450.
- Houle D, Meyer K. Estimating sampling error of evolutionary statistics based on genetic covariance matrices using maximum likelihood. *J Evol Biol*. 2015;28(8):1542–1549.
- Houle D, Morikawa B, Lynch M. Comparing mutational variabilities. *Genetics*. 1996;143(3):1467–1483.
- Houle D, Nuzhdin SV. Mutation accumulation and the effect of *copia* insertions in *Drosophila melanogaster*. *Genet Res*. 2004;83(1):7–18.
- Huang W, Lyman RF, Lyman RA, Carbone MA, Harbison ST, Magwire MM, Mackay TF. Spontaneous mutations and the origin and maintenance of quantitative genetic variation. *eLife*. 2016;5:e14625.
- Johnson T, Barton N. Theoretical models of selection and mutation on quantitative traits. *Philos Trans R Soc Lond B Biol Sci*. 2005;360(1459):1411–1425.
- Katju V, Bergthorsson U. Old trade, new tricks: insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches. *Genome Biol Evol*. 2019;11(1):136–165.
- Katju V, Packard LB, Bu L, Keightley PD, Bergthorsson U. Fitness decline in spontaneous mutation accumulation lines of *Caenorhabditis elegans* with varying effective population sizes. *Evolution*. 2015;69(1):104–116.
- Kavanaugh CM, Shaw RG. The contribution of spontaneous mutation to variation in environmental responses of *Arabidopsis thaliana*: responses to light. *Evolution*. 2005;59(2):266–275.
- Keightley PD, Mackay TF, Caballero A. Accounting for bias in estimates of the rate of polygenic mutation. *Proc R Soc Lond B Biol Sci*. 1993;253:291–296.
- Kimura M. *The Neutral Theory of Molecular Evolution*. New York (NY): Cambridge University Press; 1983.
- Klein TW. Heritability and genetic correlation: statistical power, population comparisons and sample size. *Behav Genet*. 1974;4(2):171–189.
- Klein TW, DeFries JC, Finkbeiner CT. Heritability and genetic correlation: standard errors of estimates and sample size. *Behav Genet*. 1973;3(4):355–364.
- Kondrashov AS, Houle D. Genotype-environment interactions and the estimation of the genomic mutation rate in *Drosophila melanogaster*. *Proc R Soc Lond B Biol Sci*. 1994;258:221–227.
- Littell R, Milliken G, Stroup W, Wolfinger R, Schabenberger O. *SAS for Mixed Models*. Cary (NC): SAS Institute; 2006.
- Luijckx P, Ho EKH, Stanic A, Agrawal AF. Mutation accumulation in populations of varying size: large effect mutations cause most mutational decline in the rotifer *Brachionus calyciflorus* under UV-C radiation. *J Evol Biol*. 2018;31(6):924–932.
- Lynch M. The rate of polygenic mutation. *Genet Res*. 1988;51(2):137–148.
- Lynch M. Evolution of the mutation rate. *Trends Genet*. 2010;26(8):345–352.
- Lynch M, Blanchard J, Houle D, Kibota T, Schultz S, Vassilieva L, Willis J. Perspective: spontaneous deleterious mutation. *Evolution*. 1999;53(3):645–663.
- Lynch M, Hill WG. Phenotypic evolution by neutral mutation. *Evolution*. 1986;40(5):915–935.
- Lynch M, Walsh B. *Genetics and Analysis of Quantitative Traits*. Sunderland (MA): Sinauer; 1998.
- Mackay T, Lyman RF, Hill WG. Polygenic mutation in *Drosophila melanogaster*: non-linear divergence among unselected strains. *Genetics*. 1995;139(2):849–859.
- Mackay TFC, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet*. 2009;10(8):565–577.
- Martin G, Lenormand T. The fitness effect of mutations across environments: a survey in light of fitness landscape models. *Evolution*. 2006;60(12):2413–2427.
- McGuigan K, Blows MW. Joint allelic effects on fitness and metric traits. *Evolution*. 2013;67(4):1131–1142.
- McGuigan K, Petfield D, Blows MW. Reducing mutation load through sexual selection on males. *Evolution*. 2011;65(10):2816–2829.
- Merilä J, Sheldon BC. Genetic architecture of fitness and nonfitness traits: empirical patterns and development of ideas. *Heredity*. 1999;83(2):103–109.
- Meyer K, Houle D. Sampling based approximation of confidence intervals for functions of genetic covariance matrices. *Proc Assoc Adv Anim Breed Genet*. 2013;20:523–526.
- Mezey JG, Houle D, Nuzhdin SV. Naturally segregating quantitative trait loci affecting wing shape of *Drosophila melanogaster*. *Genetics*. 2005;169(4):2101–2113.
- Mukai T. Genetic structure of natural populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability. *Genetics*. 1964;50:1–19.
- Ness RW, Morgan AD, Colegrave N, Keightley PD. Estimate of the spontaneous mutation rate in *Chlamydomonas reinhardtii*. *Genetics*. 2012;192(4):1447–1454.
- Neto-Silva RM, Wells BS, Johnston LA. Mechanisms of growth and homeostasis in the *Drosophila* wing. *Annu Rev Cell Dev Biol*. 2009;25:197–220.
- Ostrow D, Phillips N, Avalos A, Blanton D, Boggs A, Keller T, Levy L, Rosenbloom J, Baer CF. Mutational bias for body size in *Rhabditid nematodes*. *Genetics*. 2007;176(3):1653–1661.
- Reddiex AJ, Allen SL, Chenoweth SF. A genomic reference panel for *Drosophila serrata*. G3 (Bethesda). 2018;8(4):1335–1346.
- Rockman MV. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution*. 2012;66(1):1–17.
- Rohlf F. tpsRelw Version 1.45. Stony Brook: Department of Ecology and Evolution, State University of New York; 2007.
- Rohlf FJ. Shape statistics: procrustes superimpositions and tangent spaces. *J Class*. 1999;16(2):197–223.
- Rohlf FJ. tpsDig, Digitize Landmarks and Outlines, version 2.17. Stony Brook: Department of Ecology and Evolution, State University of New York; 2013.
- Schrider DR, Houle D, Lynch M, Hahn MWJG. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. *Genetics*. 2013;194(4):937–954.
- Self SG, Liang K-Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Statist Assoc*. 1987;82(398):605–610.
- Simons YB, Bullaughey K, Hudson RR, Sella G. A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biol*. 2018;16(3):e2002985.
- Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. Drift-barrier hypothesis and mutation-rate evolution. *Proc Natl Acad Sci USA*. 2012;109(45):18488–18492.
- Sztepanacz JL, Blows MW. Accounting for sampling error in genetic eigenvalues using random matrix theory. *Genetics*. 2017;206(3):1271–1284.
- van der Graaf A, Wardenaar R, Neumann DA, Tautd A, Shaw RG, Jansen RC, Schmitz RJ, Colomé-Tatché M, Johannes F. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proc Natl Acad Sci USA*. 2015;112(21):6676–6681.

Vassilieva LL, Hook AM, Lynch M. The fitness effects of spontaneous mutations in *Caenorhabditis elegans*. *Evolution*. 2000;54(4):1234–1246.

Walsh B, Lynch M. *Evolution and Selection of Quantitative Traits*. Oxford: Oxford University Press; 2018.

Wayne ML, Mackay TF. Quantitative genetics of ovariole number in *Drosophila melanogaster*. II. Mutational variation and genotype-environment interaction. *Genetics*. 1998;148(1):201–210.

Wright S. Evolution in Mendelian populations. *Genetics*. 1931;16(2):97–159.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 2010;42(7):565–569.

Communicating editor: J. Wolf