Online Supplemental S2: Additional results

for Solving the paradox of stasis: Squashed stabilizing selection and the limits of detection

Benjamin C. Haller[1] and Andrew P. Hendry[1]

1. Department of Biology and Redpath Museum, McGill University, 859 Sherbrooke Street West, Montreal, Quebec, Canada H3A 0C4

This supplemental section presents results that are ancillary or orthogonal to the focus of the main paper.

The neutral trait

The mean P(β*) for the neutral trait (a_n and z_n, taken together) was significantly less than for the selected trait (a_s and z_s, taken together), indicating that the modeled selection regime increased the rate of detection of linear selection (neutral: mean = 0.0469, SD = 0.0075, range = [0.0125, 0.0612], n = 4320, selected: mean = 0.0649, SD = 0.0576, range = [0.002, 0.577], n = 4320, one-sided paired t_{4319} = 20.96, P < 0.001). Similarly, the mean P(γ*) for the neutral trait was significantly less than for the selected trait, indicating that the modeled selection regime also increased the rate of detection of quadratic selection (neutral: mean = 0.0436, SD = 0.0107, range = [0.0056, 0.0638], n = 4320, selected: mean = 0.1365, SD = 0.2187, range = [0.000, 1.000], n = 4320, one-sided paired t_{4319} = 28.10, P < 0.001). The mean P(β*) and P(γ*) for the neutral trait are both < 0.05, indicating that selection detected on the neutral trait is likely attributable mainly to type I error.

Genetic correlation (i.e. linkage disequilibrium) between the neutral trait and the selected trait could arise due to both stochastic fluctuations and selection. To check for the existence of such correlations, linear regressions were conducted between the neutral trait values (a_n) and the selected trait values (a_s) of individuals in every generation of every realization. For each realization, the median R^2 of these regressions was calculated across all generations to determine the typical degree of correlation observed. The highest per-realization median R^2 observed was 0.022, and the median (of the per-realization median R^2 values) was only 0.0011. Most realizations (94%) had a median R^2 of less than 0.005. Although small, these correlations were often significant; the same realization with the largest median R^2 also had the smallest per-realization median P-value observed, 2.2×10^{-6}, and a significant correlation was detected in the majority of generations of 11% of realizations.

These results show that correlation certainly existed in some realizations. At the same time, the strength of these correlations was so small that the neutral and selected traits were for practical purposes uncorrelated; linkage disequilibrium of this magnitude did not substantially drive the evolution of the neutral trait. Supposing that it did, however, such an effect would in any case be conservative for testing our hypothesis: it would produce a pattern of correlated selection on the neutral trait that would increase the rate of detection of selection on it, and thus decrease the observed distinctiveness of the pattern of selection detected on the selected trait.

The neutral trait therefore appears to have provided a meaningful control, showing that the observed effects of model parameters on the selected trait were the result of selection. (See also *Autocorrelation and reproducibility* for an analysis of temporal autocorrelation of selection on the neutral trait.)

ANOVA tables for main paper analyses

ANOVA tables are here presented for the analyses given in the main paper (Tables S2.1–S2.8). Effect sizes are given as η^2, not partial η^2, calculated as the sum of squares for the effect of interest divided by the total sum of squares (Levine and Hullett 2002). See Results for further information on the analyses corresponding to these tables.

The effect of the competition width, σ_c, was not included in the ANOVAs shown in Tables S2.1, S2.2, S2.5, and S2.6, since that parameter is not used in the model when competition is not enabled.

Table S2.1. Results from ANOVA (main effects only), testing for effects of parameters on the rate of detection of linear selection, P(β*), without competition. Terms in bold are of large effect ($\eta^2 \geq 0.03$); all are also highly significant ($P < 0.001$).

	Df	SS	*F*-ratio	*P*-value	η^2
V_E	2	0.039	6.36	0.002	0.003
ω	**1**	**0.905**	**293.01**	**< 0.001**	**0.071**
m	**2**	**3.425**	**554.31**	**< 0.001**	**0.269**
T	**1**	**0.597**	**193.28**	**< 0.001**	**0.047**
N_s	2	0.283	45.82	< 0.001	0.022
G	2	0.104	16.81	< 0.001	0.008
α	1	0.156	50.56	< 0.001	0.012
μ	**1**	**0.575**	**186.21**	**< 0.001**	**0.045**
Residuals	2147	6.632			0.522

Table S2.3. Results from ANOVA (main effects only), testing for effects of parameters on the rate of detection of linear selection, P(β*), with competition. Terms in bold are of large effect ($\eta^2 \geq 0.03$); all are also highly significant ($P < 0.001$).

	Df	SS	*F*-ratio	*P*-value	η^2
V_E	2	0.000	2.43	0.088	0.002
ω	1	0.000	2.70	0.100	0.001
σ_c	1	0.000	2.31	0.129	0.001
m	2	0.007	35.72	< 0.001	0.024
T	**1**	**0.020**	**212.76**	**< 0.001**	**0.073**
N_s	2	0.007	38.43	< 0.001	0.026
G	**2**	**0.011**	**58.21**	**< 0.001**	**0.040**
α	1	0.001	13.45	< 0.001	0.005
μ	**1**	**0.026**	**278.65**	**< 0.001**	**0.095**
Residuals	2146	0.200			0.734

Table S2.5. Results from ANOVA (main effects only), testing for effects of parameters on the rate of detection of quadratic selection, P(γ*), without competition. Terms in bold are of large effect ($\eta^2 \geq 0.03$); all are also highly significant ($P < 0.001$).

	Df	SS	*F*-ratio	*P*-value	η^2
V_E	**2**	**6.284**	**90.09**	**< 0.001**	**0.053**
ω	**1**	**15.725**	**495.91**	**< 0.001**	**0.132**
m	**2**	**10.003**	**157.72**	**< 0.001**	**0.084**
T	**1**	**9.114**	**287.44**	**< 0.001**	**0.077**
N_s	**2**	**3.714**	**58.57**	**< 0.001**	**0.031**
G	2	1.662	26.21	< 0.001	0.014
α	1	0.632	19.92	< 0.001	0.005
μ	1	3.468	109.31	< 0.001	0.029
Residuals	2147	68.079			0.574

Table S2.2. Results from ANOVA (including two-way interactions), testing for effects of parameters on the rate of detection of linear selection, P(β*), without competition. Terms in bold are of large effect ($\eta^2 \geq 0.03$); all are also highly significant ($P < 0.001$).

	Df	SS	*F*-ratio	*P*-value	η^2
V_E	2	0.039	15.54	< 0.001	0.003
ω	**1**	**0.905**	**715.99**	**< 0.001**	**0.071**
m	**2**	**3.425**	**1354.50**	**< 0.001**	**0.269**
T	**1**	**0.597**	**472.29**	**< 0.001**	**0.047**
N_s	2	0.283	111.96	< 0.001	0.022
G	2	0.104	41.07	< 0.001	0.008
α	1	0.156	123.54	< 0.001	0.012
μ	**1**	**0.575**	**455.02**	**< 0.001**	**0.045**
V_E*ω	2	0.011	4.35	0.013	0.001
V_E*m	4	0.039	7.75	< 0.001	0.003
V_E*T	2	0.205	80.94	< 0.001	0.016
V_E*N_s	4	0.006	1.09	0.359	0.000
V_E*G	4	0.024	4.77	< 0.001	0.002
V_E*α	2	0.019	7.33	< 0.001	0.001
V_E*μ	2	0.066	26.01	< 0.001	0.005
ω*m	**2**	**0.963**	**381.04**	**< 0.001**	**0.076**
ω*T	1	0.297	235.05	< 0.001	0.023
ω*N_s	2	0.027	10.64	< 0.001	0.002
ω*G	2	0.001	0.53	0.591	0.000
ω*α	1	0.017	13.07	< 0.001	0.001
ω*μ	1	0.062	49.17	< 0.001	0.005
m*T	**2**	**0.496**	**196.31**	**< 0.001**	**0.039**
m*N_s	**4**	**0.385**	**76.12**	**< 0.001**	**0.030**
mG	4	0.224	44.29	< 0.001	0.018
mα	2	0.216	85.37	< 0.001	0.017
m*μ	**2**	**0.510**	**201.84**	**< 0.001**	**0.040**
T*N_s	2	0.025	9.86	< 0.001	0.002
T*G	2	0.012	4.72	< 0.001	0.001
T*α	1	0.006	4.40	0.036	0.000
T*μ	1	0.064	50.54	< 0.001	0.005
N_s*G	4	0.010	2.05	0.085	0.001
N_s*α	2	0.024	9.64	< 0.001	0.002
N_s*μ	2	0.033	12.94	< 0.001	0.003
G*α	1	0.003	2.42	0.120	0.000
G*μ	2	0.146	57.77	< 0.001	0.011
α*μ	1	0.105	82.81	< 0.001	0.008
Residuals	2086	2.637			0.207

Table S2.4. Results from ANOVA (including two-way interactions), testing for effects of parameters on the rate of detection of linear selection, P(β*), with competition. Terms in bold are of large effect ($\eta^2 \geq 0.03$); all are also highly significant ($P < 0.001$).

	Df	SS	*F*-ratio	*P*-value	η^2
V_E	2	0.000	4.36	0.013	0.002
ω	1	0.000	4.84	0.028	0.001
σ_c	1	0.000	4.14	0.042	0.001
m	2	0.007	64.02	< 0.001	0.024
T	**1**	**0.020**	**381.27**	**< 0.001**	**0.073**
N_s	2	0.007	68.86	< 0.001	0.026
G	**2**	**0.011**	**104.31**	**< 0.001**	**0.040**
α	1	0.001	24.10	< 0.001	0.005
$\boldsymbol{\mu}$	**1**	**0.026**	**499.36**	**< 0.001**	**0.095**
V_E*ω	2	0.000	3.91	0.020	0.001
V_E*σ_c	2	0.001	6.14	0.002	0.002
V_E*m	4	0.000	0.77	0.545	0.001
V_E*T	2	0.001	9.88	< 0.001	0.004
V_E*N_s	4	0.000	0.31	0.873	0.000
V_E*G	4	0.001	3.97	0.003	0.003
V_E*α	2	0.000	0.04	0.957	0.000
V_E*μ	2	0.001	5.58	0.004	0.002
$\boldsymbol{\omega*\sigma_c}$	**1**	**0.017**	**327.45**	**< 0.001**	**0.062**
ω*m	2	0.002	15.68	< 0.001	0.006
$\boldsymbol{\omega*T}$	**1**	**0.009**	**177.50**	**< 0.001**	**0.034**
ω*N_s	2	0.000	0.96	0.381	0.000
$\boldsymbol{\omega*G}$	**2**	**0.010**	**92.97**	**< 0.001**	**0.035**
ω*α	1	0.001	11.13	< 0.001	0.002
ω*μ	1	0.003	67.00	< 0.001	0.013
σ_c*m	2	0.002	21.29	< 0.001	0.008
σ_c*T	1	0.004	80.06	< 0.001	0.015
σ_c*N_s	2	0.000	0.04	0.957	0.000
σ_c*G	2	0.000	3.43	0.032	0.001
σ_c*α	1	0.001	16.20	< 0.001	0.003
σ_c*μ	1	0.003	48.50	< 0.001	0.009
m*T	2	0.001	9.16	< 0.001	0.003
$\boldsymbol{m*N_s}$	**4**	**0.008**	**39.93**	**< 0.001**	**0.030**
m*G	4	0.001	7.12	< 0.001	0.005
m*α	2	0.001	8.01	< 0.001	0.003
m*μ	2	0.000	2.17	0.114	0.001
T*N_s	2	0.000	0.40	0.669	0.000
T*G	2	0.001	13.39	< 0.001	0.005
T*α	1	0.000	0.01	0.923	0.000
$\boldsymbol{T*\mu}$	**1**	**0.015**	**284.23**	**< 0.001**	**0.054**
N_s*G	4	0.001	3.18	0.013	0.002
N_s*α	2	0.000	4.49	0.011	0.002
N_s*μ	2	0.000	1.48	0.227	0.001
G*α	1	0.001	11.24	< 0.001	0.002
G*μ	2	0.006	58.82	< 0.001	0.022
α*μ	1	0.001	12.59	< 0.001	0.002
Residuals	2073	0.108			0.395

Table S2.6. Results from ANOVA (including two-way interactions), testing for effects of parameters on the rate of detection of quadratic selection, P(γ*), without competition. Terms in bold are of large effect ($\eta^2 \geq 0.03$); all are also highly significant ($P < 0.001$).

	Df	SS	*F*-ratio	*P*-value	η^2
$\boldsymbol{V_E}$	**2**	**6.284**	**256.08**	**< 0.001**	**0.053**
$\boldsymbol{\omega}$	**1**	**15.725**	**1281.55**	**< 0.001**	**0.132**
$\boldsymbol{m}$	**2**	**10.003**	**407.60**	**< 0.001**	**0.084**
$\boldsymbol{T}$	**1**	**9.114**	**742.82**	**< 0.001**	**0.077**
$\boldsymbol{N_s}$	**2**	**3.714**	**151.35**	**< 0.001**	**0.031**
G	2	1.662	67.74	< 0.001	0.014
α	1	0.632	51.49	< 0.001	0.005
μ	1	3.468	282.64	< 0.001	0.029
$\boldsymbol{V_E*\omega}$	**2**	**5.354**	**218.19**	**< 0.001**	**0.045**
V_E*m	4	0.878	17.88	< 0.001	0.007
$\boldsymbol{V_E*T}$	**2**	**7.885**	**321.33**	**< 0.001**	**0.066**
V_E*N_s	4	0.485	9.89	< 0.001	0.004
V_E*G	4	0.108	2.20	0.066	0.001
V_E*α	2	0.042	1.72	0.179	0.000
V_E*μ	2	0.211	8.60	< 0.001	0.002
$\boldsymbol{\omega*m}$	**2**	**5.089**	**207.39**	**< 0.001**	**0.043**
$\boldsymbol{\omega*T}$	**1**	**6.936**	**565.28**	**< 0.001**	**0.058**
ω*N_s	2	2.260	92.10	< 0.001	0.019
ω*G	2	0.619	25.23	< 0.001	0.005
ω*α	1	0.246	20.04	< 0.001	0.002
ω*μ	1	1.311	106.82	< 0.001	0.011
m*T	2	2.009	81.86	< 0.001	0.017
m*N_s	4	0.823	16.78	< 0.001	0.007
m*G	4	1.015	20.68	< 0.001	0.009
m*α	2	0.407	16.60	< 0.001	0.003
m*μ	2	1.643	66.96	< 0.001	0.014
T*N_s	2	0.915	37.29	< 0.001	0.008
T*G	2	0.008	0.33	0.718	0.000
T*α	1	0.018	1.48	0.224	0.000
T*μ	1	0.200	16.26	< 0.001	0.002
N_s*G	4	0.374	7.61	< 0.001	0.003
N_s*α	2	0.149	6.07	0.002	0.001
N_s*μ	2	0.552	22.51	< 0.001	0.005
G*α	1	0.193	15.76	< 0.001	0.002
G*μ	2	2.120	86.39	< 0.001	0.018
α*μ	1	0.631	51.43	< 0.001	0.005
Residuals	2086	25.595			0.216

Table S2.8. Results from ANOVA (including two-way interactions), testing for effects of parameters on the rate of detection of quadratic selection, P(γ*), with competition. Terms in bold are of large effect ($\eta^2 \geq 0.03$); all are also highly significant ($P < 0.001$).

	Df	SS	*F*-ratio	*P*-value	η^2
V_E	2	0.116	4.53	0.011	0.001
ω	**1**	**9.339**	**729.50**	**< 0.001**	**0.107**
σ_c	**1**	**6.665**	**520.62**	**< 0.001**	**0.076**
m	2	2.574	100.53	< 0.001	0.029
T	1	0.319	24.93	< 0.001	0.004
N_s	**2**	**5.135**	**200.55**	**< 0.001**	**0.059**
G	**2**	**5.069**	**197.98**	**< 0.001**	**0.058**
α	1	0.417	32.61	< 0.001	0.005
μ	1	0.026	2.00	0.158	0.000
V_E*ω	2	0.212	8.30	< 0.001	0.002
V_E*σ_c	2	0.222	8.67	< 0.001	0.003
V_E*m	4	0.045	0.89	0.471	0.001
V_E*T	2	0.309	12.05	< 0.001	0.004
V_E*N_s	4	0.095	1.86	0.114	0.001
V_E*G	4	0.004	0.07	0.991	0.000
V_E*α	2	0.002	0.07	0.931	0.000
V_E*μ	2	0.013	0.49	0.610	0.000
ω*σ_c	**1**	**8.886**	**694.10**	**< 0.001**	**0.102**
ω*m	2	1.330	51.94	< 0.001	0.015
ω*T	1	0.127	9.92	0.002	0.001
ω*N_s	2	2.583	100.90	< 0.001	0.030
ω*G	**2**	**3.734**	**145.85**	**< 0.001**	**0.043**
ω*α	1	0.531	41.50	< 0.001	0.006
ω*μ	1	0.452	35.28	< 0.001	0.005
σ_c*m	2	0.754	29.44	< 0.001	0.009
σ_c*T	1	0.090	7.03	0.008	0.001
σ_c*N_s	2	1.936	75.61	< 0.001	0.022
σ_c*G	**2**	**4.461**	**174.24**	**< 0.001**	**0.051**
σ_c*α	1	0.559	43.69	< 0.001	0.006
σ_c*μ	1	0.495	38.68	< 0.001	0.006
m*T	2	0.011	0.42	0.655	0.000
m*N_s	4	0.716	13.98	< 0.001	0.008
m*G	4	0.570	11.13	< 0.001	0.007
m*α	2	0.031	1.21	0.299	0.000
m*μ	2	0.017	0.68	0.508	0.000
T*N_s	2	0.065	2.52	0.081	0.001
T*G	2	0.024	0.93	0.397	0.000
T*α	1	0.000	0.00	0.962	0.000
T*μ	1	0.011	0.87	0.350	0.000
N_s*G	4	1.796	35.07	< 0.001	0.021
N_s*α	2	0.111	4.35	0.013	0.001
N_s*μ	2	0.022	0.87	0.417	0.000
G*α	1	0.428	33.45	< 0.001	0.005
G*μ	2	0.552	21.54	< 0.001	0.006
α*μ	1	0.138	10.76	0.001	0.002
Residuals	2073	26.538			0.303

Table S2.7. Results from ANOVA (main effects only), testing for effects of parameters on the rate of detection of quadratic selection, P(γ*), with competition. Terms in bold are of large effect ($\eta^2 \geq 0.03$); all are also highly significant ($P < 0.001$).

	Df	SS	*F*-ratio	*P*-value	η^2
V_E	2	0.116	2.15	0.117	0.001
ω	**1**	**9.339**	**346.31**	**< 0.001**	**0.107**
σ_c	**1**	**6.665**	**247.15**	**< 0.001**	**0.076**
m	2	2.574	47.73	< 0.001	0.029
T	1	0.319	11.84	< 0.001	0.004
N_s	**2**	**5.135**	**95.21**	**< 0.001**	**0.059**
G	**2**	**5.069**	**93.99**	**< 0.001**	**0.058**
α	1	0.417	15.48	< 0.001	0.005
μ	1	0.026	0.95	0.330	0.000
Residuals	2146	57.870			0.661

The strength of stabilizing selection

Values of 1.0 and 10.0 were used for ω, the width of the stabilizing fitness function (Table 1). These values are difficult to interpret, however, since the magnitude of the "unit of ecological phenotype" is defined only relative to the magnitudes of other model parameters also based upon that unit (V_E, α, and, with competition, σ_c; Table 1). It is thus more informative to consider the strength of selection relative to the variation in the trait under selection. Following Estes and Arnold (2007), we can standardize ω by squaring it and dividing by the population phenotypic standard deviation of the selected trait (also expressed in the model's units of ecological phenotype) to obtain a standard interpretable metric that we will here call ω_s^2. This standardization is complicated slightly by the fact that the population's phenotypic variance is variable in our model, and changes to fit the selective regime imposed by the parameters of each realization. Once a realization has reached a dynamic equilibrium (after the "burn-in" period, in other words), the phenotypic variance is relatively stable, however, so we can estimate ω_s^2 for each realization using the median of the per-generation phenotypic variances of that realization.

Some generations (~4%) have a phenotypic standard deviation of zero because there is no variation in the selected trait in that generation; this occurs due to the elimination of alternative genotypes by selection or drift. These generations

would produce an estimate of infinity for ω_s^2; these generations are thus excluded from the analysis. Generations with a very small but non-zero phenotypic standard deviation produce very large estimates of ω_s^2 for the same reason; we will thus report the median, first quartile, and third quartile of ω_s^2 estimates to omit these uninformative outliers.

Following this procedure for all of our core (non-supplementary) realizations, we observed a median ω_s^2 of 17.5 (Q1 = 5.1, Q3 = 91.7). The strongest standardized selection observed across these realizations was 2.20; the weakest was 125,000 (due to a very small phenotypic standard deviation; see above). These values agree with the range of empirical estimates of the standardized strength of stabilizing selection (Estes and Arnold 2007, their Fig. 7), which range from close to zero to somewhat above 100, but are typically less than 50 and show a strong mode at about 3; our observed distribution of ω_s^2 also has a strong mode at 3, and indeed, closely resembles the empirical distribution (not shown).

Because the phenotypic variance of the population responds strongly to the presence of negative frequency-dependent selection, ω_s^2 depends not only on ω, but also on whether competition is enabled in the model (Table S2.9).

Table S2.9. Effects of the width of the fitness function, ω, and the presence or absence of competition, C, on the estimated standardized strength of stabilizing selection, ω_s^2 (first quartile Q1, median, and third quartile Q3). Smaller ω_s^2 values represent stronger selection in standardized units.

ω	C	Q1 ω_s^2	median ω_s^2	Q3 ω_s^2
all	all	5.1	17.5	91.7
1.0	NO	5.6	17.9	63.3
1.0	YES	3.2	9.1	31.8
10.0	NO	31.2	99.0	313.8
10.0	YES	3.9	6.1	12.7

In particular, competition broadens the phenotypic distribution and thus increases the standardized strength of selection, from a median of 17.9 to 9.1 for $\omega = 1.0$, and from a median of 99.0 to 6.1 for $\omega = 10$ (Table S2.9). Interestingly, this effect of negative frequency-dependent selection in the context of squashed stabilizing selection can apparently be so pronounced that weaker stabilizing selection in absolute terms produces stronger stabilizing selection in standardized units; in Table S2.9, for example, the median ω_s^2 with competition is 9.1 under strong stabilizing selection ($\omega = 1$), but strengthens to 6.1 under weak stabilizing selection ($\omega = 10$). This illustrates the remarkable extent to which the population phenotypic variance adjusts to the selective regime imposed.

Effects of mutational variance

The genetic architecture G, mutation rate μ, and mutation effect size α all had an effect upon the rate of detection of selection (Figs. 3 and 4, Tables S2.1–S2.8). However, these three factors together determine the mutational variance V_M (Table S1.1), and so V_M may more parsimoniously explain some or all of their effects. The effects of V_M are shown in Fig. S2.1.

The rate of detection of directional selection, P(β*), increased strongly with increasing V_M when competition was off (linear regression, $t_{2158} = 12.20$, $P < 0.001$, $\beta = 65.2$), explaining a significant portion of the variance among realizations (adj. $R^2 = 0.0640$, $F(1, 2158) = 148.7$, $P < 0.001$; Fig. S2.1a). With competition, there was very little variation in P(β*), and while a significant effect of V_M was observed (linear regression, $t_{2158} = 5.94$, $P < 0.001$, $\beta = 4.8$), it explained little variance (adj. $R^2 = 0.0156$, $F(1, 2158) = 35.2$, $P < 0.001$; not shown).

The rate of detection of quadratic selection, P(γ*), increased strongly with increasing V_M without competition (linear regression, $t_{2158} = 11.66$, $P < 0.001$, $\beta = 190.8$) and explained significant variance among realizations (adj. $R^2 = 0.0588$, $F(1, 2158) = 135.9$, $P < 0.001$) (Fig. S2.1b). With competition, a moderate positive association was observed (linear regression, $t_{2158} = 5.40$, $P < 0.001$, $\beta = 77.8$), although it explained relatively little variance (adj. $R^2 = 0.0129$, $F(1, 2158) = 29.18$, $P < 0.001$) (Fig. S2.1c).

High levels of mutational variance thus appeared to drive higher rates of detection of selection in many realizations, but the size of this effect was not generally large, and both linear and quadratic selection were detected at levels substantially exceeding the type I error rate for many realizations even at the lowest levels of mutational variance modeled. Our conclusions therefore appear robust to

the level of mutational variance, although of course the quantitative frequency with which selection is detected changes.

Considering G, μ and α together as V_M may not always be valid, of course, as these parameters may have independent effects. Their independent effects were observed to be fairly small in most cases (Tables S2.1–S2.8), although the effect of mutation rate μ on the rate of detection of linear selection was substantial, particularly with competition (Tables S2.1–S2.4); this is unsurprising, since the more frequently mutations occur, the more frequently selection against maladapted mutants with extreme phenotypes may be expected, producing detectable directional selection. Details of genetic architecture did not generally affect the qualitative interpretation of the effects of other parameters (minor interaction effects, and the direction of main effects, are noted in Results). This is in accord with the conclusion of Slatkin (1979) that genetic details should not substantially affect the equilibrium reached under stabilizing and frequency-dependent selection, as long as the genetics impose no constraints upon the mean or variance of the trait; our triallelic model did impose some such constraints, but they were minimal, by design (see Supplemental S1, *Genetic architectures*).

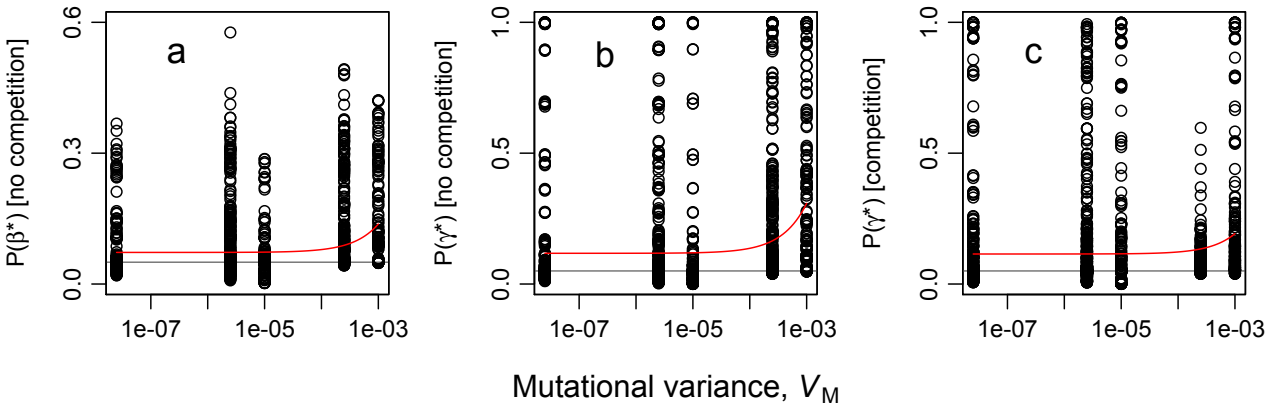

Figure S2.1. The effect of mutational variance, V_M, on the rate of detection of: (a) linear selection, P(β*), without competition (realizations with competition not shown; see text); (b) quadratic selection, P(γ*), without competition; (c) quadratic selection, P(γ*), with competition. Red curves show significant linear regression fits (curved due to the logarithmic x-axis scale). Gray lines show the expected type I error rate of 0.05. Note difference in vertical scales.

Effects of heritability

The narrow-sense heritability, h^2, in this model is not specified as a parameter; rather, it is an emergent property. The heritability was calculated for each generation of each realization, for both the selected trait (using a_s and z_s) and the neutral trait (using a_n and z_n), as shown by

$$h^2 = V_A / V_P \qquad \text{(Formula S2.1; Falconer 1989),}$$

where V_A is the additive genetic variance and V_P is the phenotypic variance. For each realization, the mean heritability across all generations was then calculated, and that was the basis for further analyses; hereafter "heritability" and h^2 refer to these per-realization mean values.

The observed heritability was positively correlated with mutational variance, V_M, and negatively correlated with environmental variance, V_E (Fig. S2.2a), as expected. The heritability also depended strongly on the evolutionary dynamics of the realization, so it exhibited a much wider range of values for the selected trait than for the neutral trait, for given values of V_M and V_E (Fig. S2.2b). Indeed, all three genetic architectures showed a heritability close to zero for the neutral trait at the lowest level of mutational variance, because fixation due to drift was common (Fig. S2.2a), but this was not as much of a problem for the selected trait (Fig. S2.2b), particularly with negative frequency-dependent selection to balance the fitnesses of multiple alleles.

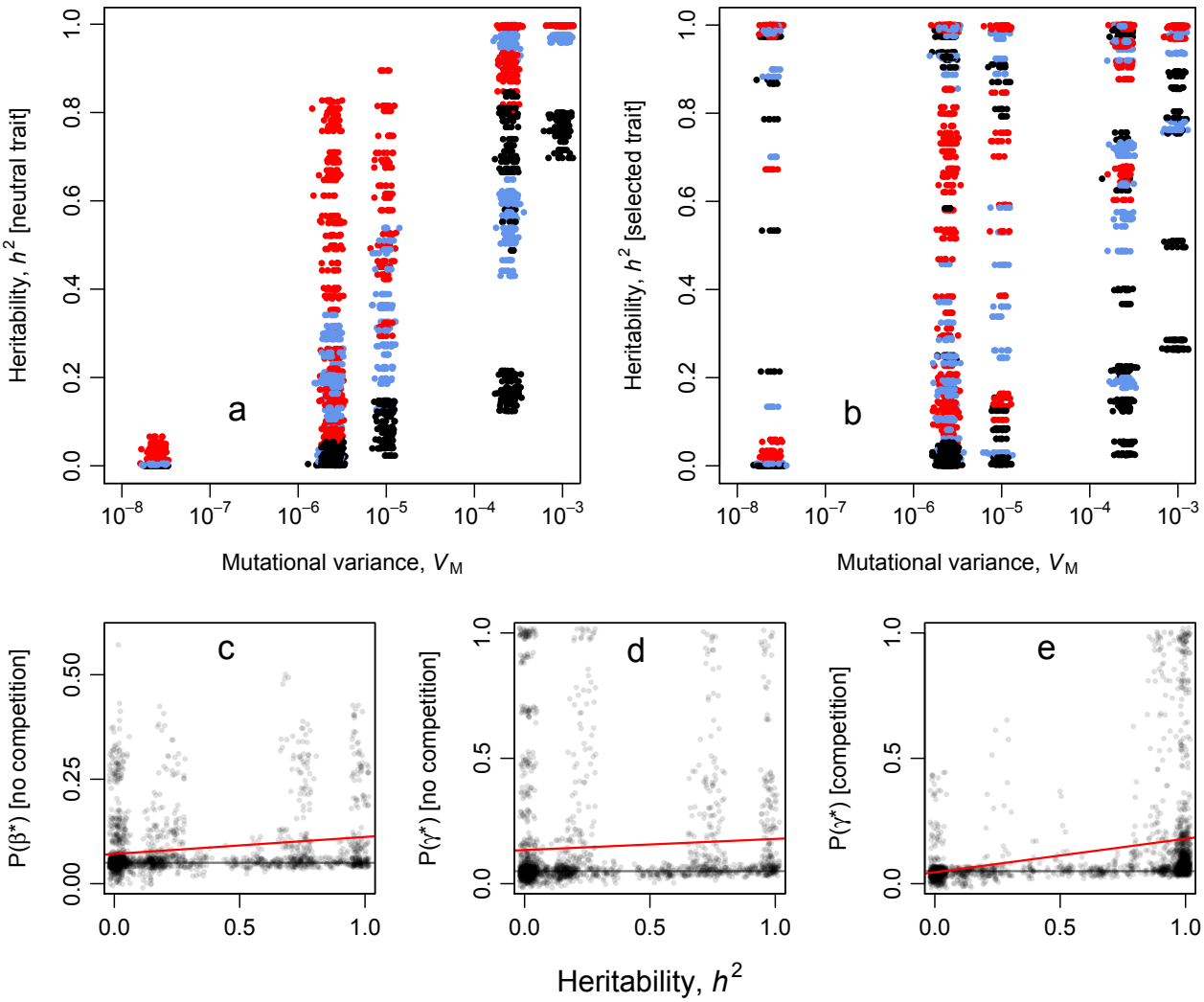

Figure S2.2. Origins and effects of heritability. Panels (a) and (b) show the effects of mutational variance, V_M, and environmental variance, V_E, on heritability of: (a) the neutral trait; and (b) the selected trait. Color indicates the value of V_E (red: 0.001, blue: 0.01, black: 0.1). Jitter was added on the *x*-axis to separate the points. Panels (c)–(e) show the effect of heritability on the rate of detection of: (c) linear selection, P(β*), without competition (realizations with competition not shown; see text); (d) quadratic selection, P(γ*), without competition; and (e) quadratic selection, P(γ*), with competition. Red lines show significant linear regression fits. Gray lines show the expected type I error rate of 0.05. Transparency and a small amount of jitter were used to better show the density of points. Note difference in vertical scales.

The heritability of a trait is a measure of the amount of additive genetic variation present in a population, relative to the total variation present (Formula S2.1); as such, it is often interpreted as representing the "evolvability" or "variability" of a trait (Houle 1992). This is convenient because heritabilities are unitless, and can thus be compared across traits and even across systems; however, this interpretation of heritability can be problematic, and standardization of the additive genetic variance using the trait mean instead of the phenotypic variance has been recommended (Houle 1992; Hansen and Houle 2008; Hansen et al. 2011). Unfortunately, there are several cases in which the mean-standardized standard deviation, also called the coefficient of variation or CV, cannot be used. One such case is for "interval-scale" rather than "ratio-scale" traits, for which the zero point is arbitrary and has no special meaning; the evolvability of interval-scale traits must be

compared using heritabilities, for want of a better alternative (Houle 1992; Hansen and Houle 2008; Hansen et al. 2011).

Although it should be interpreted with caution, then, we can examine whether the heritabilities observed in our model for the selected trait correspond to the distribution of heritabilities observed in nature. Mousseau and Roff (1987) show that across a large number of empirical studies the median heritability is ~0.3, with a standard deviation of ~0.25. Across all realizations, the median heritability we observed was 0.23 (SD = 0.42), somewhat lower than the empirical value, but with higher variability such that the full range of empirically observed heritabilities was included (indeed, we observed heritabilities spanning 0.0 to 1.0, depending upon model parameters). The heritability depended strongly upon the presence of competition (without competition, median = 0.08, SD = 0.35; with competition, median = 0.90, SD = 0.43); this is expected since negative frequency-dependent selection promotes the retention of genetic variation, which was otherwise lost completely in some generations. Heritability also depended strongly on the genetic details of the mutation rate, mutational effect size, and genetic architecture; however, since our conclusions were robust to variation in those parameters (see Results, and *Effects of mutational variance*), the effects of those parameters on heritability appear to ultimately be of little consequence for the qualitative pattern of detection of selection.

We can examine this more closely by looking directly at the correlation between heritability and the detection of selection. Heritability had mixed effects on the rate of detection of selection (Figs. S2.2c–e). For linear selection, heritability significantly increased that rate without competition (linear regression, $t_{2158} = 8.77$, $P < 0.001$, $\beta = 0.0407$, adj. $R^2 = 0.034$; Fig. S2.2c), but with competition heritability had a much smaller effect, although it explained more variability (linear regression, $t_{2158} = 14.21$, $P < 0.001$, $\beta = 0.0076$, adj. $R^2 = 0.0852$; not shown). For quadratic selection, heritability's effect without competition was significantly positive but explained little variance (linear regression, $t_{2158} = 3.08$, $P = 0.002$, $\beta = 0.0444$, adj. $R^2 = 0.0039$; Fig. S2.2d), whereas with competition the positive effect was somewhat stronger and more explanatory (linear regression, $t_{2158} = 13.91$, $P < 0.001$, $\beta = 0.133$, adj. $R^2 = 0.0818$; Fig. S2.2e).

In no case did heritability explain even 10% of the variation in the rate of detection of selection. In summary, then, the heritabilities observed in our model were roughly congruent with the heritabilities observed in empirical studies, and in any case heritability appeared to have only a small effect upon the detection of selection.

The low importance of heritability for the detection of selection is not unexpected given our other results. In particular, for selection to be detected, the most important factors should be the strength of selection actually exerted on the population, and the amount of phenotypic variation present for that selection to act upon. The proportion of variation present that is due to additive genetics – the heritability – is essential for predicting the magnitude of the *response* to selection in the next generation, but it is irrelevant to the detection of selection itself. Indeed, selection should be readily detectable on traits with no genetic basis whatsoever, as long as phenotypic variation (due to developmental stochasticity or environmental influences) is still present in sufficient quantity. The question that began this section – does our model exhibit a realistic amount of genetically-based variation? – is thus not highly relevant to our conclusions. More important is the question: does our model exhibit a realistic amount of phenotypic variation relative to the strength of selection imposed upon the population? We have shown above that it does (see *The strength of stabilizing selection*).

Selective deaths and the detection of selection

The design of our model allows the effect of deaths during the random mortality phase (Figs. 3d, 4d) to be separated from the effect of deaths due to selection ("selective deaths"). A higher mean selective death rate was positively associated with the rate of detection of linear selection, P(β*), without competition (Kendall's $\tau = 0.298$, $P < 0.001$; Spearman's $\varrho = 0.433$, $P < 0.001$; Fig. S2.3a). A positive association also existed with competition, but it was quite weak (Kendall's $\tau = 0.072$, $P < 0.001$; Spearman's $\varrho = 0.123$, $P < 0.001$; not shown). The selective death rate was positively

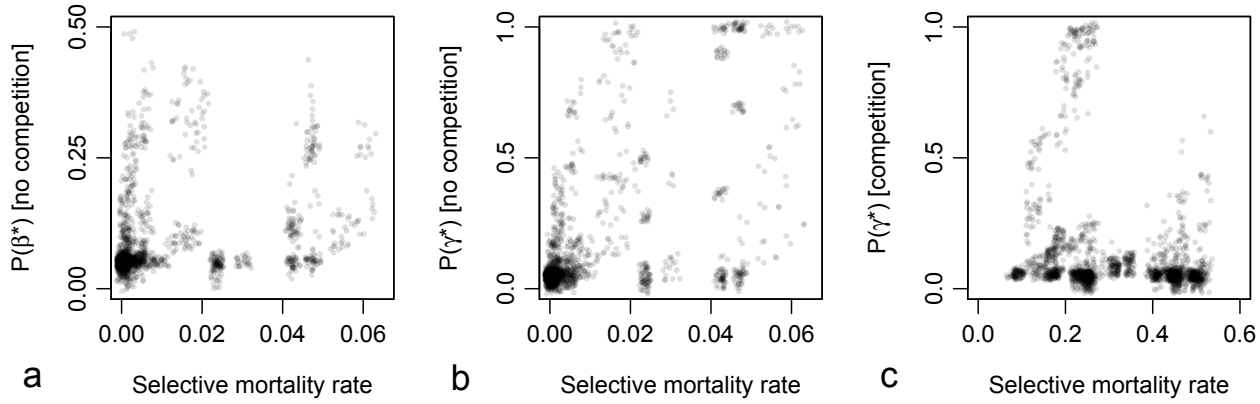

Figure S2.3. Effects of the selective death rate (rate of selective death per generation) on the rate of detection of: (a) linear selection, P(β*), without competition (realizations with competition not shown; see text); (b) quadratic selection, P(γ*), without competition; (c) quadratic selection, P(γ*), with competition. Transparency and a small amount of jitter were used to better show the density of points. The association in (a) and (b) is positive, while that in (c) is negative (see text).

associated with the rate of detection of quadratic selection, P(γ*), without competition (Kendall's $\tau = 0.352$, $P < 0.001$; Spearman's $\varrho = 0.490$, $P < 0.001$; Fig. S2.3b). With competition, however, the association was negative; the greater the number of selective deaths, the lower was the rate of detection of quadratic selection (Kendall's $\tau = -0.202$, $P < 0.001$; Spearman's $\varrho = -0.307$, $P < 0.001$; Fig. S2.3c). For further illustration of patterns of selective death and the detection of selection, see *Two case studies*.

The detection of selection, whether linear or quadratic, is essentially a problem of detecting a signal (a meaningful pattern of selective deaths) amid a lot of noise (random mortality, plus stochasticity in the selective deaths). Many of our results can be understood from this perspective, as particular factors boosted or attenuated the signal, or increased or decreased the noise masking that signal. Intuitively, the strength of the signal should be based upon the number of selective deaths observed, and this was often the case: the higher the selective death rate was, the stronger the signal was, and so the more likely it was that selection was detected (Figs. S2.3a–b). Interestingly, however, this was not the case with squashed stabilizing selection (Fig. S2.3c). In this case, selective deaths occurred across the whole phenotypic range of the population; deaths of extreme phenotypes were "stabilizing", while deaths close to the mean phenotype were "disruptive", and both types typically occurred in every generation. Differentiating between stabilizing and disruptive selection means determining whether extreme phenotypes are less fit or more fit than the mean phenotype; but under squashed stabilizing selection, the population was trapped at a dynamic equilibrium at which all phenotypes, whether extreme or average, had a low fitness due to the flattened shape of the fitness function. The higher the selective death rate, the more consistent this mixed signal was; adding more signal – more selective deaths – thus actually decreased the rate of detection of quadratic selection.

This signal-to-noise perspective is also helpful for understanding the effect of random mortality, since it affects the amount of stochastic noise masking the signal of the selective deaths. As postulated, random mortality made detection of selection more difficult for both linear and quadratic selection (Figs. 3d, 4d), both with and without competition (Tables S2.1–S2.8); even 10% random mortality per generation made the detection of selection quite unlikely, particularly without competition. The number of selective deaths per generation was quite low without competition, and so the signal was easily drowned out by even a small amount of noise. With competition there were many more selective deaths, but the pattern of those deaths was relatively unhelpful for detecting selection, as discussed above, and so the signal was often still balanced at the edge of detectability.

Finally, there is the question of the trait analyzed: the genetic breeding value, a_s, or the phenotype, z_s. For detection of linear selection, this had a small (although significant) effect on results (Fig. 3e). Quadratic selection, however, was much more likely to be detected using phenotypic values rather than genotypic values, especially without competition (Fig. 4e, Tables S2.5–S2.8). This seems logical, since the phenotype is the thing upon which selection actually acts; the correlation between phenotype and survival ought to be stronger than the correlation between genotype and survival. In this sense, using genetic values actually adds noise to the analysis; subtracting out the "noise" of environmental variance actually adds noise that obscures the correlation that we are trying to detect.

Effects of parameters on the selection gradient distribution

The Results section *Distribution of selection gradient values* shows average distributions across all the realizations conducted, separated only by whether competition was on or off; but these distributions also depended upon the other independent variables. The effects of all independent variables upon the distributions of β and γ are shown in Figs. S2.4–S2.12.

No attempt is made here to assess statistically whether the differences between these distributions are significant or not; each histogram is based on an extremely large number of gradient estimates (typically $> 3.5\times10^7$), so almost any differences visible to the eye are likely to be significant and reproducible. For example, since the competition width, σ_c, is not used by the model when competition is turned off, Figs. S2.6a and S2.6e are actually independent replicates, but they are essentially indistinguishable; likewise for Figs. S2.6b and S2.6f. Additionally, the level of symmetry apparent in all of the linear gradient distributions gives an indication of the reproducibility of these distributions.

However, each histogram is a composite of all realizations of the model with a specified value of the independent variable (and with or without competition, as the case might be). This produces artifacts, because it means that each histogram shows the combined effects produced by the particular discrete parameter values used in our realizations. For example, close examination of Fig. S2.9a reveals that the distribution of significant β estimates is pentamodal: it has five distinct peaks. This is not an indication that the quantitative genetic architecture produces a pentamodal distribution of β estimates; it is merely the visible consequence of particular peak positions being produced by the particular parameter values used in our realizations. The best way to read these histograms, therefore, is to compare them within columns, to see how the different values of the independent variable affect things such as the width of the distribution, the frequency of detection of selection, the prevalence of positive versus negative values, the likelihood of significant estimates near zero, and so forth.

This artifact from combining histograms generated with discrete parameter values has one particularly striking manifestation: the sharp, high central peaks of the observed selection gradient distributions. For example, when analysis is based on the genetic trait value, a_s, the distribution of selection gradients tends to be narrower, with a higher central peak (Figs. S2.8a–d), compared to analysis based on the phenotypic trait value, z_s, which typically produces a broader distribution with a lower central peak (Figs. S2.8e–h). Histograms that combine results from both the genetic and phenotypic analyses (all histograms except Fig. S2.8, that is) thus tend to have a broader shape but a higher central peak, as a result of the combination of these two distributional shapes. In fact, the same effect occurs whenever results are combined from model parameterizations that resulted in different distributional breadths, notably different values of m (Fig. S2.7), N_s (Fig. S2.9), and μ (Fig. 2.12). This effect is responsible for much of the double-exponential (Laplace) shape of all of the presented histograms; the distribution of β and γ estimates within a single realization tends to be roughly Gaussian (not shown), but the combination of these Gaussian distributions with different breadths results in a leptokurtic distribution. The same effect is likely responsible for the leptokurtic shapes observed in the distributions of selection gradients in the meta-analyses of, e.g., Kingsolver et al. (2001); those leptokurtic shapes might be the result of combining β or γ estimates from many different natural systems. Each system might exhibit a roughly

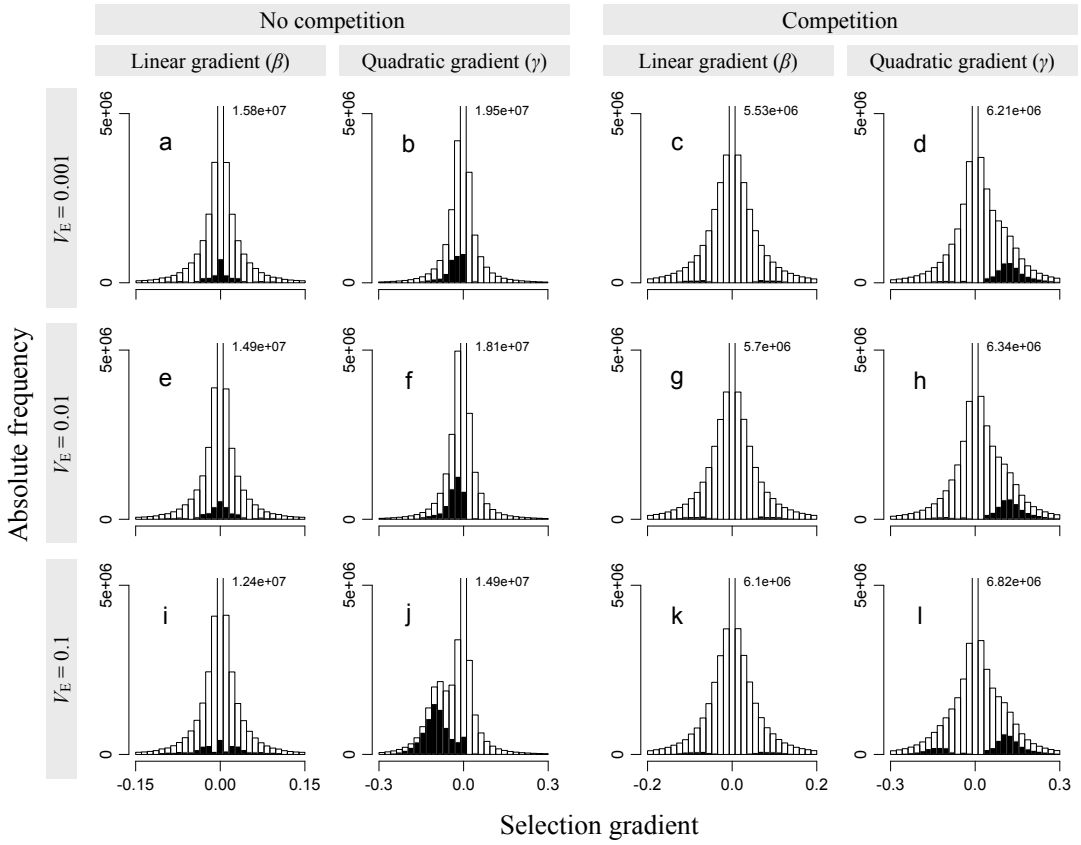

Figure S2.4. Effects of environmental variance, V_E, on the distribution of estimates of β and γ. In all panels, black shading indicates those estimates that are significant ($P < 0.05$). Note x-axis scales are only guaranteed to match within columns. The heights of peaks extending beyond the plot are labeled.

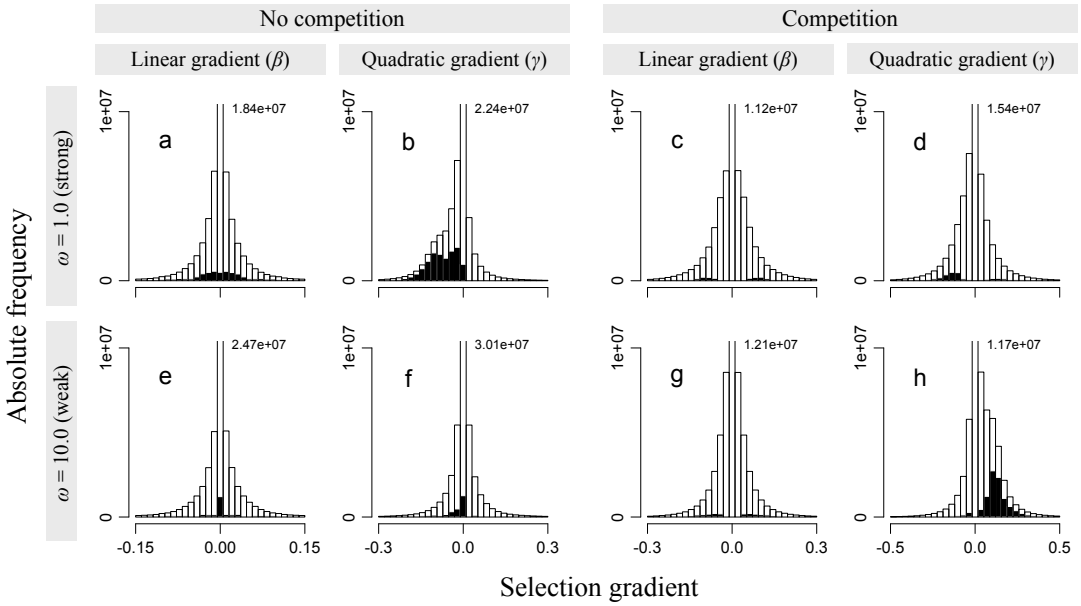

Figure S2.5. Effects of the fitness function width, ω, on the distribution of estimates of β and γ. In all panels, black shading indicates those estimates that are significant ($P < 0.05$). Note x-axis scales are only guaranteed to match within columns. The heights of peaks extending beyond the plot are labeled.

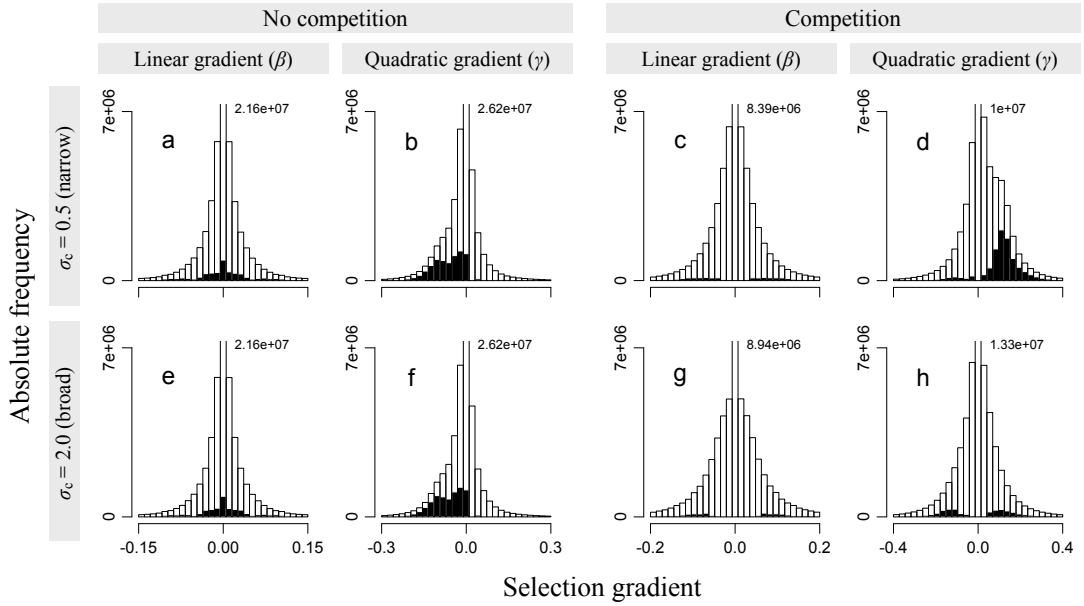

Figure S2.6. Effects of competition width, σ_c, on the distribution of estimates of β and γ. In all panels, black shading indicates those estimates that are significant ($P < 0.05$). Note x-axis scales are only guaranteed to match within columns. The heights of peaks extending beyond the plot are labeled.

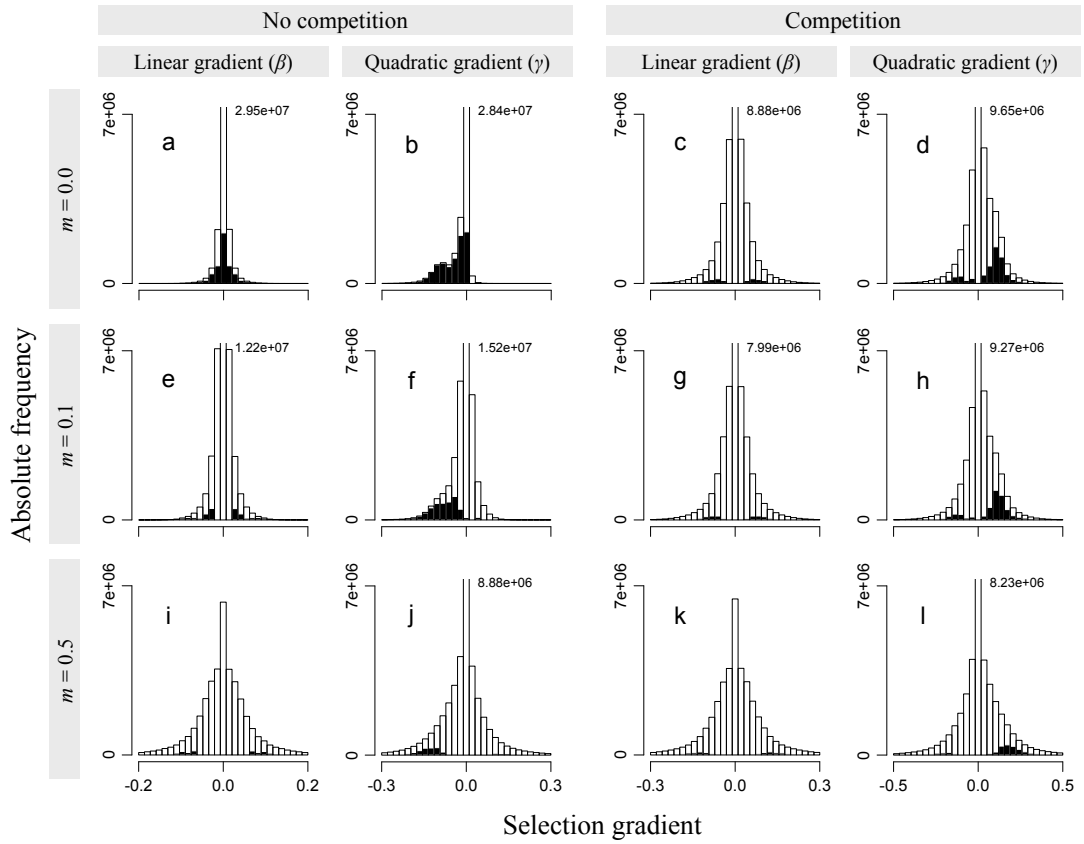

Figure S2.7. Effects of mortality rate, m, on the distribution of estimates of β and γ. In all panels, black shading indicates those estimates that are significant ($P < 0.05$). Note x-axis scales are only guaranteed to match within columns. The heights of peaks extending beyond the plot are labeled.

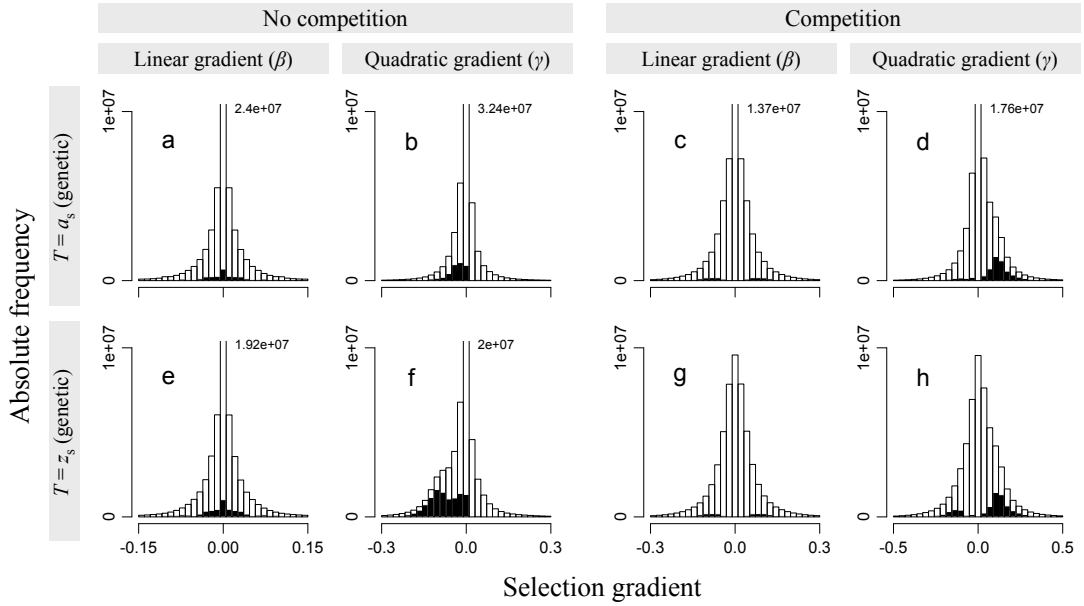

Figure S2.8. Effects of the trait (genetic or phenotypic) examined, T, on the distribution of estimates of β and γ. In all panels, black shading indicates those estimates that are significant ($P < 0.05$). Note x-axis scales are only guaranteed to match within columns. The heights of peaks extending beyond the plot are labeled.

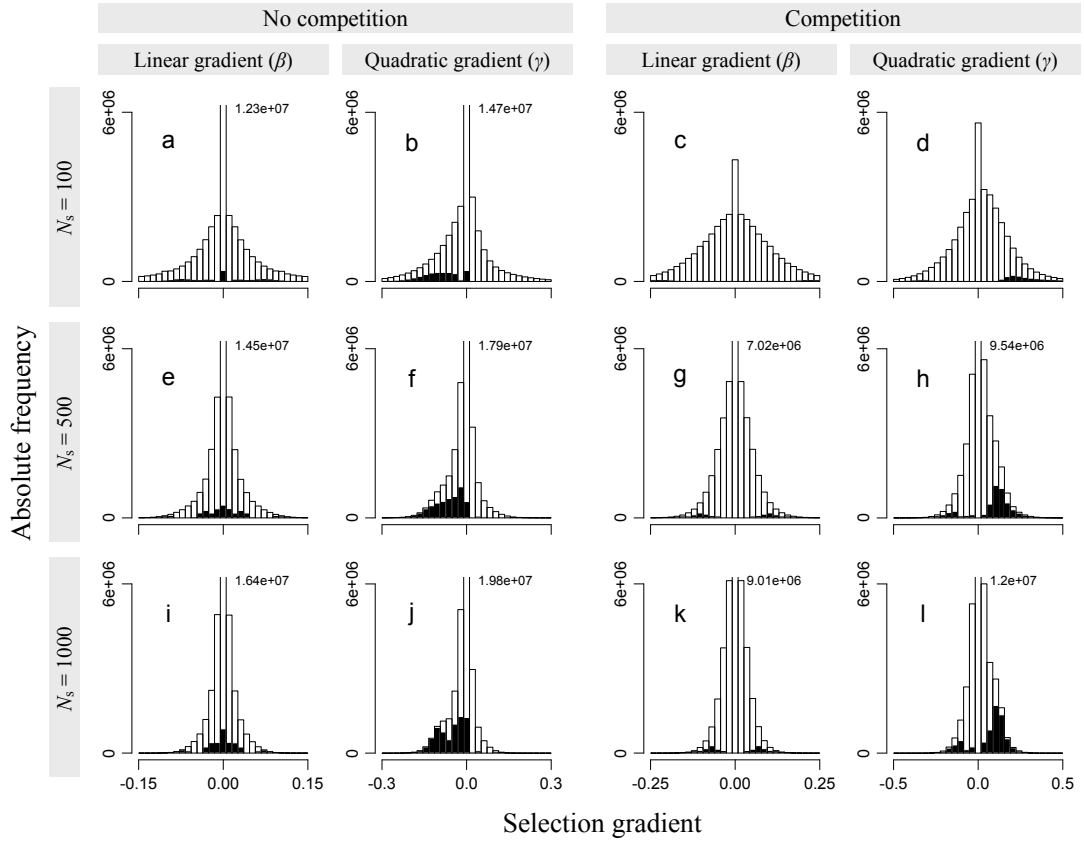

Figure S2.9. Effects of sample size, N_s, on the distribution of estimates of β and γ. In all panels, black shading indicates those estimates that are significant ($P < 0.05$). Note x-axis scales are only guaranteed to match within columns. The heights of peaks extending beyond the plot are labeled.

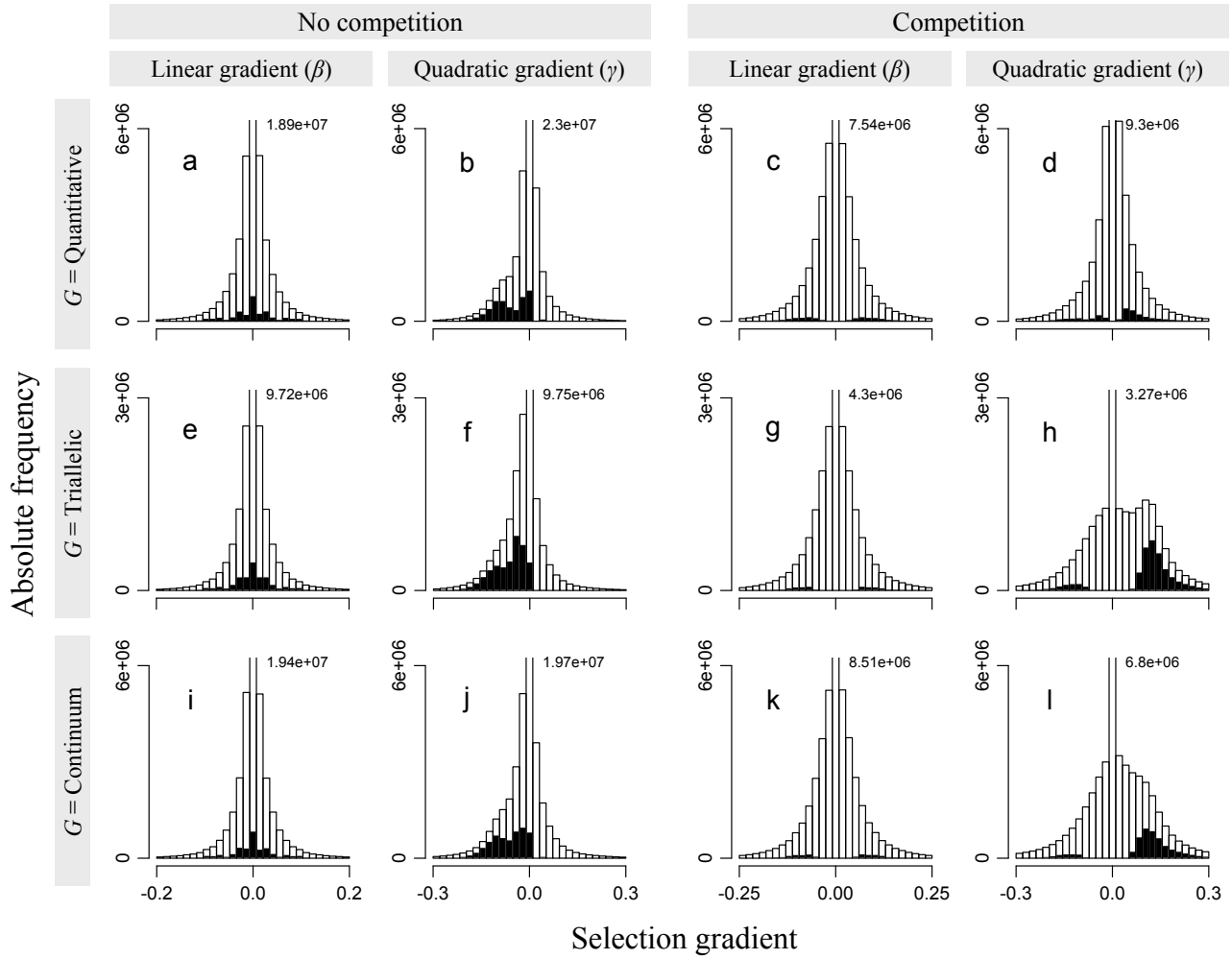

Figure S2.10. Effects of genetic architecture, G, on the distribution of estimates of β and γ. The y-axis scale for the triallelic architecture (middle row) is half that of the other architectures, to compensate for the fact that half as many realizations were conducted for that architecture (because α is not used by that architecture). In all panels, black shading indicates those estimates that are significant ($P < 0.05$). Note x-axis scales are only guaranteed to match within columns. The heights of peaks extending beyond the plot are labeled.

Gaussian distribution of selection gradient estimates, but each might have a different distributional breadth (whether due to different strengths of selection, or to different magnitudes of sampling error, across systems), such that the combination of estimates across systems yields the characteristic leptokurtic distributional shape also observed in our results.

Finally, it should be noted that considering only significant regression results can lead to a large bias in the selection gradient values considered, particularly if most of the significant regressions are the result of type I error. Caution should be used, therefore, in interpreting the patterns of significant selection in these figures, particularly in cases in which significance is rarely found.

Nevertheless, many of the patterns discussed in other sections can be clearly seen here as shifts in the distributions of β and γ. The increase in detection of stabilizing selection with a narrow fitness function and no competition, for example, can be seen in Fig. S2.5b vs. S2.5f, while the increase in detection of disruptive selection with a broad fitness function and competition can be seen in Fig. S2.5d vs. S2.5h (see also Figure 6, and Results, *Distribution of selection gradient values*). The effect of high rates of random mortality in obscuring the detection of selection is clearly seen in Fig. S2.7,

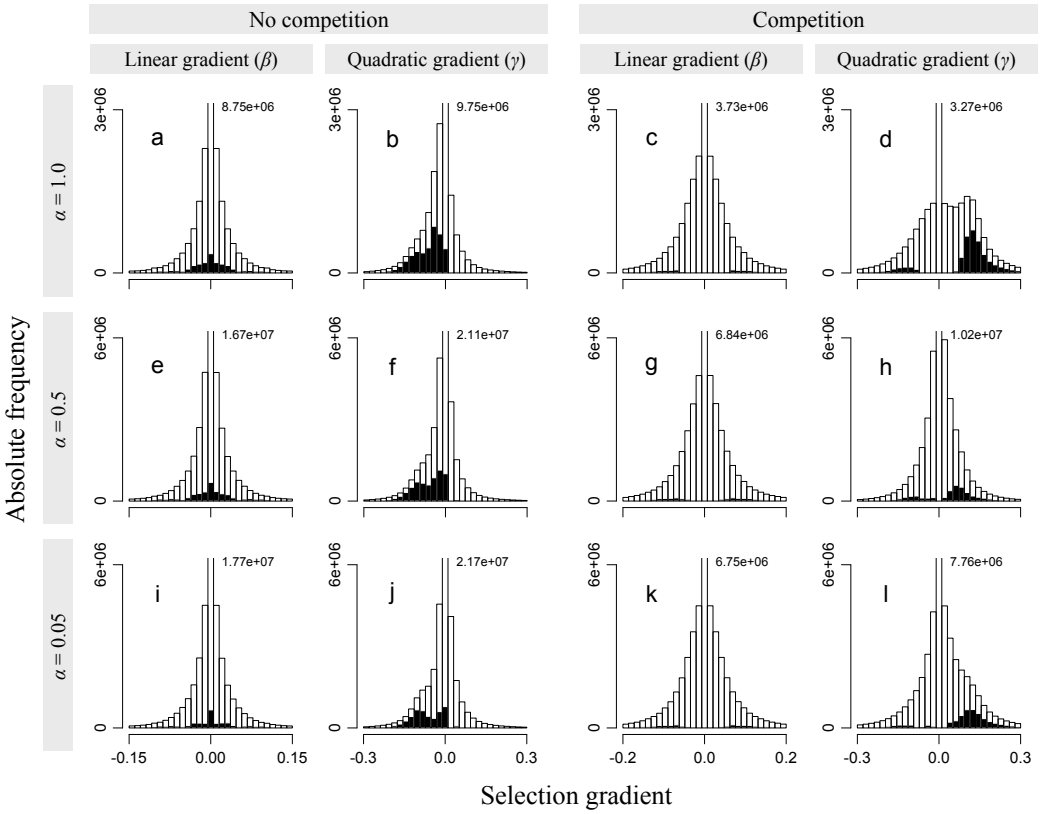

Figure S2.11. Effects of mutational effect size, α, on the distribution of estimates of β and γ. Note that $\alpha = 1.0$ only applies to the triallelic genetic architecture, while $\alpha = 0.5$ and $\alpha = 0.05$ only apply to the quantitative and continuum architectures; see Supplemental S1, *Genetic architectures*. For this reason, there are half as many realizations with $\alpha = 1.0$; the *y*-axis of that row has been adjusted accordingly. In all panels, black shading indicates those estimates that are significant ($P < 0.05$). Note *x*-axis scales are only guaranteed to match within columns. The heights of peaks extending beyond the plot are labeled.

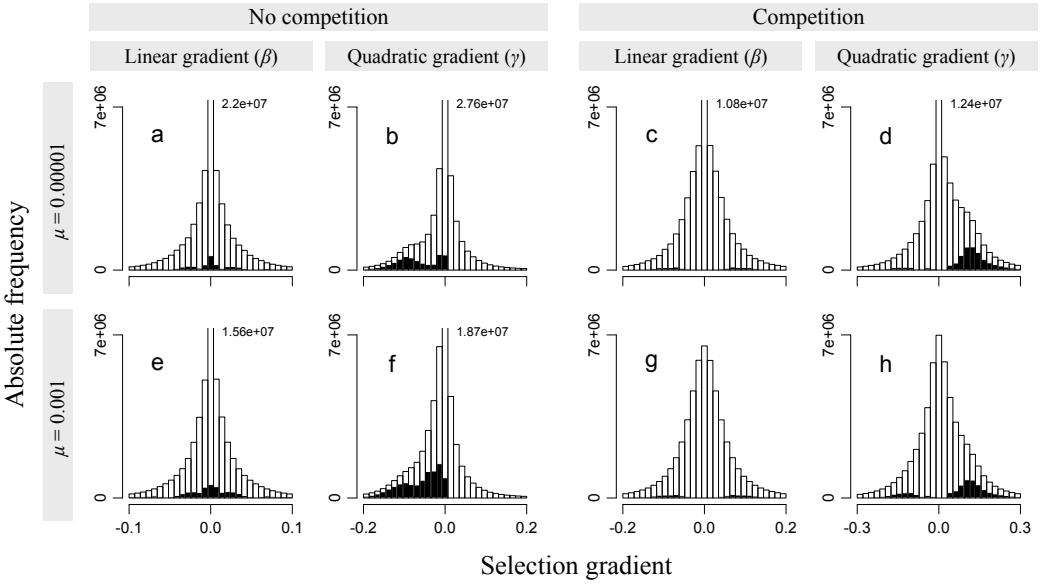

Figure S2.12. Effects of mutation rate, μ, on the distribution of estimates of β and γ. In all panels, black shading indicates those estimates that are significant ($P < 0.05$). Note *x*-axis scales are only guaranteed to match within columns. The heights of peaks extending beyond the plot are labeled.

whereas the opposing effect of large sample size can be seen in Fig. S2.9.

Two case studies

Other results presented are composites across many realizations of the model, showing average effects and general trends. We here present more specific results from two realizations of the model, one without competition (Fig. S2.13), one with competition (Fig. S2.14). The parameter values for these realizations were chosen to provide relatively strong "signal" (high rates of selective death and detection of selection), for purposes of illustration.

The mean phenotype wandered considerably over time, both without and with competition (Figs. S2.13, S2.14; note the difference in the *x*-axis scales). Without competition, there were few selective deaths, they tended to occur principally when the mean was far from the optimum, and the detection of selection followed this pattern (Fig. S2.13). With competition, there were a great many selective deaths (and so only 500 generations are shown, for clarity), and the detection of selection appeared to be largely uncorrelated with the wandering of the phenotypic mean (Fig. S2.14).

Fig. S2.13 suggests that the detection of selection without competition might have been temporally autocorrelated; detection of selection might have occured in "runs." Testing this formally, without competition significant autocorrelation was found in the significance (0 or 1) of both β (Ljung–Box $Q = 45.3$, $P < 0.001$) and γ (Ljung–Box $Q = 50.4$, $P < 0.001$), and persisted for approximately 150 generations (Figs. S2.15a–b). With competition, autocorrelation was not significant for the significance of either β (Ljung–Box $Q = 0.48$, $P = 0.49$) or γ (Ljung–Box $Q = 2.1$, $P = 0.15$), as is apparent visually (Figs. S2.15c–d).

Autocorrelation was also observed in the β and γ estimates themselves (without regard for the significance of those estimates). Without competition, significant autocorrelation was found in the estimates of both β (Ljung–Box $Q = 182.8$, $P < 0.001$) and γ (Ljung–Box $Q = 20.7$, $P < 0.001$); for β this autocorrelation persisted for ~7000 generations (Fig. S2.16a, not fully shown), while for γ it persisted for ~150 generations (Fig. S2.16b). With competition, autocorrelation was significant for the estimates of both β (Ljung–Box $Q = 145.5$, $P < 0.001$) and γ (Ljung–Box $Q = 24.2$, $P < 0.001$), but this autocorrelation was short-lived, up to perhaps only ten generations (Figs. S2.16c–d).

The autocorrelation structure of these particular realizations is detailed here for purposes of illustration; a more general analysis of the autocorrelation of model realizations is presented in the next section, *Autocorrelation and reproducibility*.

Autocorrelation and reproducibility

A metric was desired for the temporal autocorrelation of selection, following Siepielski et al. (2009) and Morrissey and Hadfield (2012) (see *Temporal variation in selection* for another perspective on this question). Since our model is of stabilizing selection, we expected that over the long term directional selection would be equally likely to be towards smaller as towards larger values, so a metric of the proportion of generations observed with a positive or negative sign for β, such as used by Siepielski et al. (2009), would not be useful. Similarly, a metric of the rate at which directional selection acts in opposite directions in two temporal replicates, as proposed by Morrissey and Hadfield (2012), would be expected to have a value very close to 0.5 for our study, since two randomly drawn temporal replicates from a timeline of 50,000 generations would be expected to be uncorrelated. Instead, then, we calculated the autocorrelation of several metrics of selection over all of the generations of a realization, for generation lags from 1 to 25000, using the standard formula for the estimation of autocorrelation,

$$\hat{R}(k) = \frac{1}{(n-k)\sigma^2} \sum_{t=1}^{n-k} (X_t - \mu)(X_{t+k} - \mu)$$

(Formula S2.2),

where X_t is the metric value at a given generation t, n is the length of the time series (50,000 generations), k is the lag (in generations), σ^2 is the variance among metric values, and μ is the mean metric value. This formula was used to calculate the autocorrelation of the selection gradient estimates themselves (β and γ) to assess whether selection of a particular type, direction and strength tends to occur in "runs"

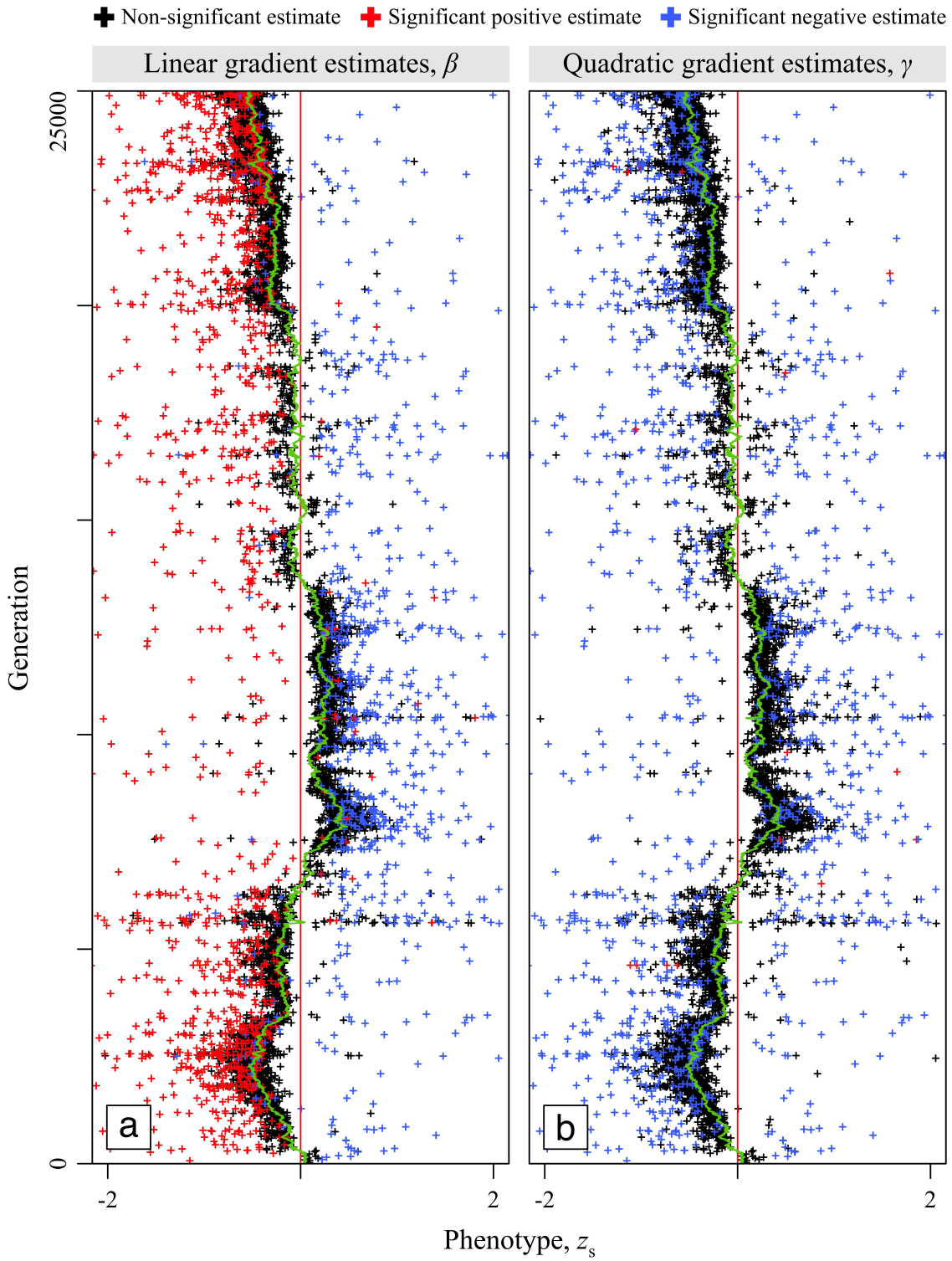

Figure S2.13. Selection over time, for a realization without competition: (a) linear selection, (b) quadratic selection. Crosses show selective deaths, colored by the significance and sign of the selection gradient, β or γ: black = not significant, red = significant and positive, blue = significant and negative. For quadratic selection, a positive gradient is consistent with disruptive selection, while a negative gradient is consistent with stabilizing selection. The central red line shows the adaptive peak, $\theta = 0$. The green line shows the population mean phenotype. Parameter values used: $m = 0$, $\omega = 10.0$, $\mu = 0.001$, $\alpha = 0.5$, $\sigma_c = 2.0$, $V_E = 0.001$, $N_j = 1000$, $N_s = 1000$, quantitative genetic architecture.

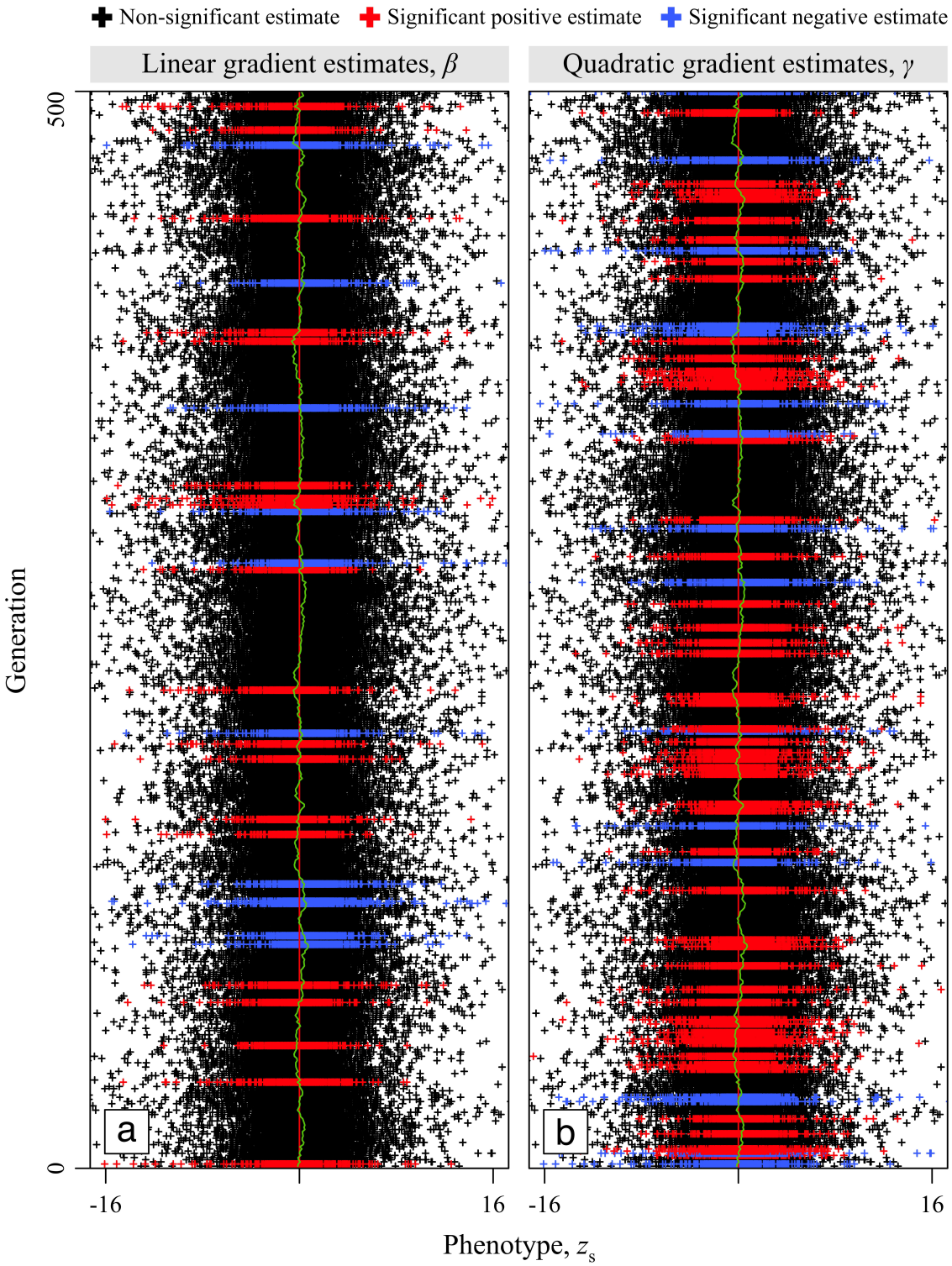

Figure S2.14. Selection over time, for a realization with competition: (a) linear selection, (b) quadratic selection. Crosses show selective deaths, colored by the significance and sign of the selection gradient, β or γ: black = not significant, red = significant and positive, blue = significant and negative. For quadratic selection, a positive gradient is consistent with disruptive selection, while a negative gradient is consistent with stabilizing selection. The central red line shows the adaptive peak, $\theta = 0$. The green line shows the population mean phenotype. Parameter values are as in Fig. S2.13 (except that competition is on); note, however, that axis scales are necessarily different than in Fig. S2.13.

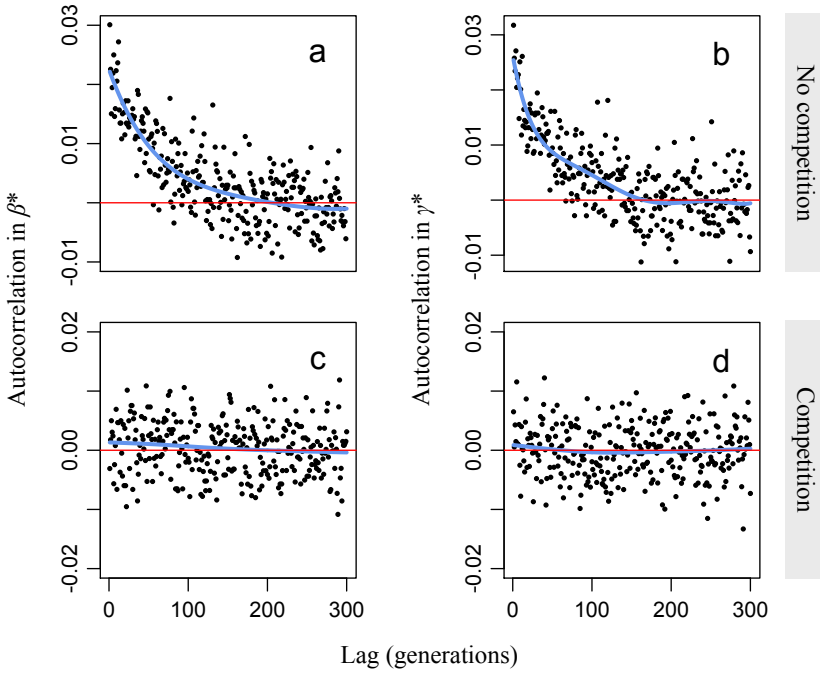

Figure S2.15. Autocorrelation of the significance of selection gradient estimates, β^* and γ^*, for the pair of runs shown in Figs. S2.13 and S2.14: (a) no competition, autocorrelation of β^*; (b) no competition, autocorrelation of γ^*; (c) competition, autocorrelation of β^*; (d) competition, autocorrelation of γ^*. Blue splines are fitted to the data for visualization. Red lines are at zero; above the red lines is the zone of positive temporal autocorrelation.

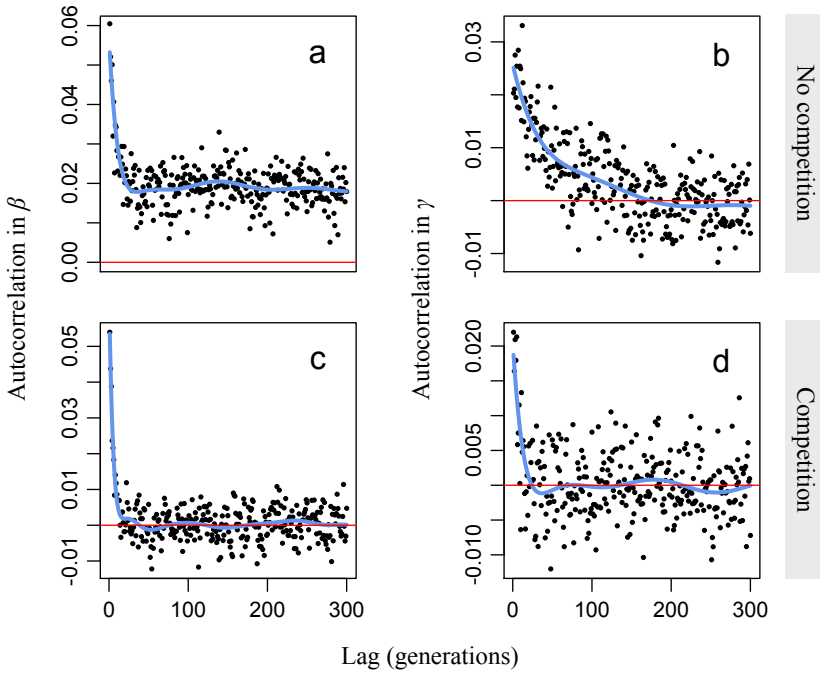

Figure S2.16. Autocorrelation in the estimated values of β and γ, for the pair of runs shown in Figs. S2.13 and S2.14: (a) no competition, autocorrelation of β; (b) no competition, autocorrelation of γ; (c) competition, autocorrelation of β; (d) competition, autocorrelation of γ. Blue splines are fitted to the data for visualization. Red lines are at zero; above the red lines is the zone of positive temporal autocorrelation.

(without regard for the significance of the selection gradient estimates). The formula was also used to calculate autocorrelation in the significance of the selection gradients (β^* and γ^*; significant = 1, non-significant = 0), to assess whether the detection of selection tends to occur in runs (without regard for the direction or strength of that selection). This analysis was only performed for the selection gradients estimated from phenotypic values.

Given autocorrelation functions as produced by Formula S2.2, several metrics were calculated to assess the significance, strength, and duration of autocorrelation. First, Ljung–Box Q tests were used to assess whether significant autocorrelation was present for a given realization. Second, the strength of short-term autocorrelation was assessed as the value of the autocorrelation function for a lag of one, $\hat{R}(1)$. Finally, the decay time until autocorrelation was no longer significant was measured as the first lag for which the autocorrelation function, averaged across a centered 25-lag-wide sliding window, was outside the 95% confidence region for significant autocorrelation as calculated using the standard formula

$$B_\alpha = \frac{z_{1-\alpha/2}}{\sqrt{N}} \qquad \text{(Formula S2.3)},$$

where z is the quantile function of the standard normal distribution, α is the significance level, and N is the sample size upon which the autocorrelation function is based. In our analysis, $\alpha = 0.05$ and $N = 50000$, so $B_\alpha \approx 0.0088$; autocorrelation values of greater absolute magnitude than this threshold are thus significant at a 95% confidence level.

Autocorrelation was never meaningful for the neutral trait. The percentage of realizations in which the Ljung–Box Q test found that autocorrelation was significant was almost exactly the significance threshold of the test (for β, 5.37% of realizations; for γ, 4.58%; for β^*, 4.44%; for γ^*, 5.23%); instances in which significant autocorrelation was found for the neutral trait are therefore likely to be type I error. The strength of short-term autocorrelation, $\hat{R}(1)$, was symmetrically distributed with a median < 0.0001, a median absolute deviation (MAD) of ≈ 0.0045, and a maximum value less than 0.021 for all four metrics (β, γ, β^*, γ^*); in other words, autocorrelation from one generation to the next for the neutral trait was equally likely to be positive or negative, and was very close to zero. Finally, the decay time for autocorrelation of the neutral trait was essentially the lowest measurable; in almost all cases, the first sliding window sampled, across lags 1 to 25, had an average autocorrelation that was non-significant, producing an estimated decay time of 13 generations. The remaining cases (0.12% of the total) appear to have been driven by a single outlier resulting from sampling noise, and showed a decay time based upon the first window position that excluded that outlier; their decay time was never greater than 43 generations. All patterns discussed below for autocorrelation of the selected trait may thus be inferred to be caused by the action of selection upon the trait, directly or indirectly.

Meaningful autocorrelation was commonly observed for the selected trait. The Ljung–Box Q test indicated significant autocorrelation at well above the type I error rate for all metrics (for β, 38.7% of realizations; for γ, 26.0%; for β^*, 10.2%; for γ^*, 12.5%). The strength of short-term autocorrelation, $\hat{R}(1)$, was strongly skewed towards positive values; medians were still close to zero (< 0.004) due to the large number of realizations for which significant autocorrelation did not exist, but MADs were somewhat larger than for the neutral trait (for β, 0.0094; for γ, 0.0068; for β^*, 0.0047; for γ^*, 0.0051), and the maximum $\hat{R}(1)$ value observed was substantially larger (for β, 0.465; for γ, 0.348; for β^*, 0.078; for γ^*, 0.068). Finally, the decay time for autocorrelation of the selected trait was sometimes quite long (for β, a maximum decay time of 16003 generations; for γ, 17458 generations; for β^*, 10481 generations; for γ^*, 14975 generations), although the median realization for all four metrics exhibited the shortest measurable decay time (13 generations), as with the neutral trait.

These results show that autocorrelation in selection gradient estimates was stronger, lasted longer, and was more detectable than autocorrelation in the significance of selection; this is to be expected, since even non-significant gradient estimates may still exhibit autocorrelation. For this reason, we now focus on the autocorrelation of the selection gradient estimates; patterns for the autocorrelation of the significance of the selection gradient estimates were qualitatively the same, but were weaker.

Realizations with the largest short-term autocorrelation, $\hat{R}(1)$, for β generally had lower random mortality m, larger sample size N_s, higher mutation rate μ, smaller ω, and were without competition. Realizations with the largest $\hat{R}(1)$ for γ showed the same trends. Realizations with the longest decay time for β generally had lower random mortality m, and were without competition. Realizations with the longest decay time for γ generally had a larger fitness function width ω, used the continuum genetic architecture, and were with competition. Even for realizations combining these parameters, however, it should be noted that the "long tail" of autocorrelation observed involved low autocorrelation values. On the other hand, it is quite surprising that any amount of autocorrelation in selection could persist for so long, given the stochasticity inherent in the dynamics of a small population.

Realizations exhibited a strong association between the autocorrelation in a selection gradient estimate and autocorrelation in the significance of that selection gradient estimate; for example, a strong association was observed between the autocorrelation of β and the autocorrelation of the significance of β. This was true for both the magnitude of short-term autocorrelation and the autocorrelation decay time (four Spearman rank correlation tests, all $\varrho \geq 0.26$; four Kendall rank correlation tests, all $\tau \geq 0.17$; all $P < 0.001$).

Realizations also exhibited associations between the magnitude of short-term autocorrelation and the length of decay time for β and γ, and between metrics for β and metrics for γ; positive associations existed between each pair of these four metrics (six Spearman rank correlation tests, all $\varrho \geq 0.22$; six Kendall rank correlation tests, all $\tau \geq 0.16$; all $P < 0.001$).

Given this observed temporal autocorrelation in both the significance of selection and selection gradient estimates, the possibility existed that temporal autocorrelation could affect other results; if many generations of a realization experienced the same selective regime, due to such autocorrelation, the statistics gathered on that realization might not be an unbiased sample. As explained in the main paper (see Methods, *Data collection*), our dataset contains two replicates of each parameter value combination without competition, because realizations were done for low and high σ_c even though that parameter was not used by the model without competition. If autocorrelation caused any bias, these replicate runs would differ substantially, because their (random) initial state or early transient dynamics would substantially condition the dynamics for the remainder of the realization.

Comparisons were thus conducted between all non-competition realizations with $\sigma_c = 0.5$ and with $\sigma_c = 2.0$, on both the neutral and the selected trait. Welch's unpaired *t*-tests were used to compare these datasets, subdivided by neutral/selected trait, against each other, testing for a difference between the means of four metrics: (1) median β, (2) median γ, (3) rate of significance of β, P(β*), and (4) rate of significance of γ, P(γ*). Each compared dataset contained 1080 observations of each metric. No test showed a significant difference at a Bonferroni-corrected α of 0.00625 (given the total of eight *t*-tests conducted), and the smallest *P*-value was > 0.04. For the metrics expected to indicate a problem with bias due to autocorrelation (the median β value and the median significant β value of the selected trait), the *P*-values were 0.168 and 0.968.

These results indicate that the results of our realizations were highly reproducible and therefore free of autocorrelation-derived bias. Nevertheless, to the extent that temporal autocorrelation might have distorted our results, it would have done so by making individual realizations more idiosyncratic, because the outcome of each realization would then be biased by its initial state and by stochastic events early in the run that echoed across the remainder of the realization. This would make patterns across multiple realizations less consistent, and thus less apparent; it would increase the variance among realizations. To the extent that our conclusions are based upon patterns of differences between means or medians measured across many realizations, then, they should be robust to problems introduced by temporal autocorrelation, since removal of any autocorrelation-caused bias would only strengthen the patterns observed, by decreasing among-realization variance.

Logistic vs. linear regression

In addition to linear regressions, logistic regressions were also performed on all realizations. All results

reported in other sections were based upon the linear regressions (except where explicitly noted), for reasons explained below. In this section we compare those results to results from the logistic regressions. In general, analyses performed using logistic regression produced qualitatively very similar results to the results based on the linear regressions; we here focus on differences.

Following standard practice, absolute fitness, rather than relative fitness, was analyzed as a function of the standardized trait values for the logistic regressions (Janzen and Stern 1998). The same eight regressions were conducted as with linear regression (linear/quadratic × genetic/phenotypic × selected/neutral) per generation per subsampled history. Regression coefficients from the logistic regressions were used to calculate an average linear selection gradient, β_{avggrad}, according to the formula

$$\beta_{\text{avggrad}} = \left[\frac{1}{N} \sum_{i=1}^{N} W(z)\left(1 - W(z)\right) \right] \alpha$$

(Formula S2.4; Janzen and Stern 1998),

where $W(z)$ is the absolute fitness of trait value z, N is the size of the population sample, and α (here) is the logistic regression coefficient for the linear term. Quadratic regression coefficients from the logistic regressions were used to calculate average quadratic selection gradients, γ_{avggrad}, using the same formula (*mutatis mutandis*), since the formula is equally applicable to this case (Janzen, F. J., and Stern, H. S., *pers. comm.*). Quadratic selection gradients were doubled, as with linear regression (Stinchcombe et al. 2008).

Logistic regression was substantially less likely to detect selection than was linear regression (Figs. S2.17a–b). For the neutral trait (a_n and z_n, taken together) the means of P(β*) and P(γ*) across all realizations were 0.0469 and 0.0436 for linear regression, marginally under the significance threshold of 0.05; for logistic regression these values were 0.0388 and 0.0302, substantially less. This trend also held for the selected trait (a_s and z_s, taken together), for which linear regression gave means of 0.0649 and 0.136, while logistic regression yielded 0.0537 and 0.113. Interestingly, the extent of this difference appeared to depend in some way upon the particular pattern of selective deaths generated by particular sets of parameters; there were many realizations in which the two methods found significance at almost identical rates, but many other realizations in which logistic regression found significance less often – sometimes much less often. This difference seemed to occur particularly when the statistical methods had few data points with which to work: without competition, and without random mortality (Figs. S2.17a–b, blue points versus other points). In such circumstances, logistic regression appeared to be very strongly conservative compared to linear regression.

Average gradients from logistic regression (β_{avggrad} and γ_{avggrad}) were markedly smaller than gradients from linear regression (β and γ), and reached a plateau relative to the linear regression gradients (Figs. S2.17c–d). The model realizations with the largest median $|\beta|$ values (~0.4) had median $|\beta_{\text{avggrad}}|$ values about four times smaller (Figs. S2.17c); similarly, the realizations with the largest median $|\gamma|$ values (~0.8) had median $|\gamma_{\text{avggrad}}|$ values approximately four times smaller (Figs. S2.17d). It is also noteworthy that increasing the sample size resulted in a decrease in the typical (median) selection gradient observed using both regression methods (Figs. S2.17c–d, point colors); this fits well with the same finding in Kingsolver et al. (2001) with respect to the empirical detection of selection, suggesting that the true strength of selection in nature is probably typically towards the weak end of the range observed in their meta-analysis (see also *Effects of parameters on the selection gradient distribution*).

For these reasons, patterns in the data were harder to see using the logistic regressions – there were fewer generations in which selection is detected, and the selection gradients found were smaller. More importantly, using linear regression for our main analysis is conservative with respect to our hypothesis that selection should be detected relatively rarely (see Introduction); using logistic regression would bias our findings in the direction of our hypothesis. For these reasons, the linear regressions were used for most results.

Substantial differences are apparent between these two methods. It is important to stress, however, that we do not attempt here to pass judgment as to which method is better. Regarding the difference in their rate of detection of selection, it may be that logistic regression failed to find

selection when it ought to have (type II error), or it may be that linear regression incorrectly found selection when it did not truly have sufficient basis (type I error) – or both methods may be correct, since they ask different statistical questions, even though both are commonly used to estimate the strength of selection. Similarly, although there was often a large difference in the magnitude of the selection gradient estimates found by the two methods, we do not judge in this analysis which method is superior. These differences should be of great concern to those who are measuring selection in the wild, since it means that results from the two methods are not comparable. For the time being, we recommend that selection should be estimated and reported using both methods. Further work on the statistical underpinnings of the estimation of selection will be important to advancing our understanding of selection, since we cannot understand what we cannot measure with both accuracy and precision.

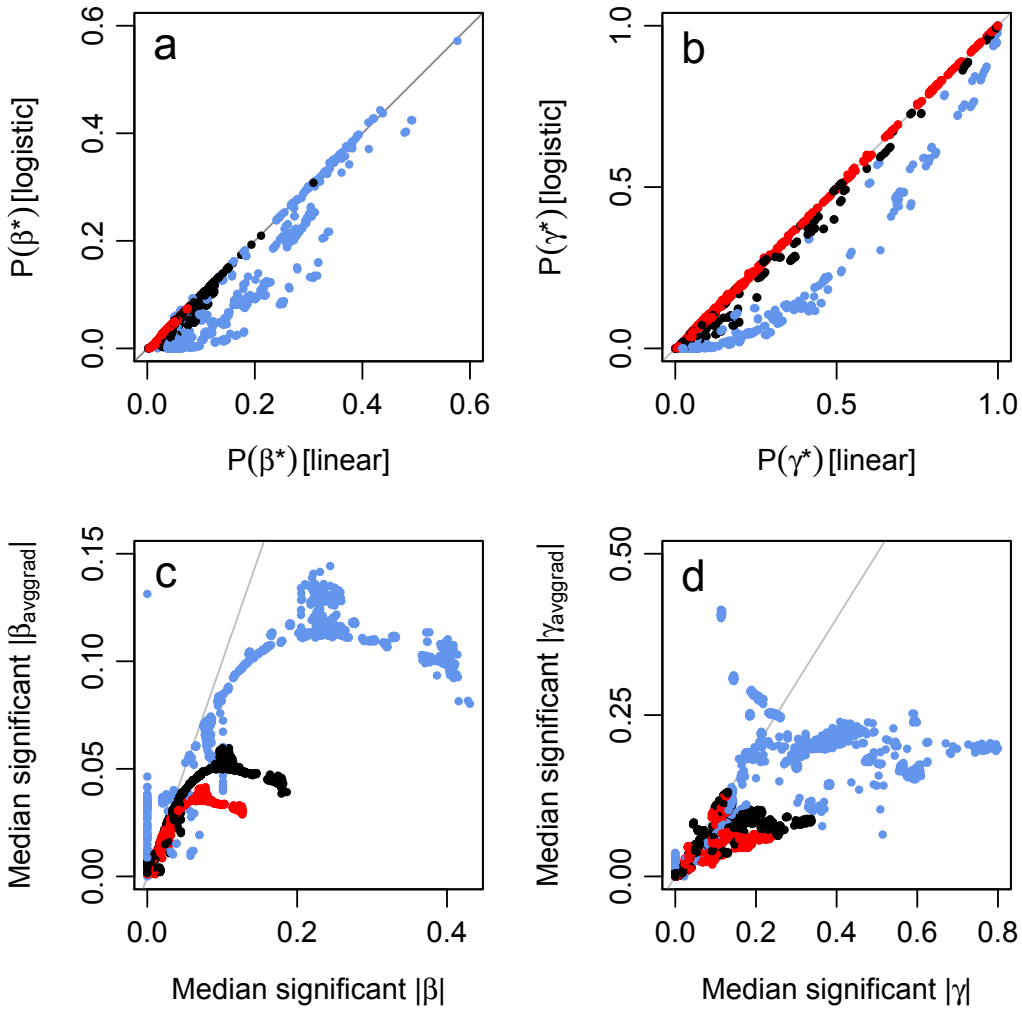

Figure S2.17. A comparison of results from linear regression (*x*-axes) and logistic regression (*y*-axes): (a) the rate of detection of linear selection, P(β*); (b) the rate of detection of quadratic selection, P(γ*); (c) the median absolute β ($\beta_{avggrad}$) across all generations for which β ($\beta_{avggrad}$) was significant; (d) the median absolute γ ($\gamma_{avggrad}$) across all generations for which γ ($\gamma_{avggrad}$) was significant. In (a) and (b), points are red if competition was on for that realization, otherwise they are blue for mortality $m = 0.0$, otherwise black. In (c) and (d), colors indicate the sample size N_s: red = 1000, black = 500, blue = 100. In all panels, the gray line is at $x = y$; note that in (c) and (d) the *x*- and *y*-axes have different scales. Panel (d) omits 3.0% of points because they have very large *y*-axis values (up to ~7000), to allow the bulk of the data to be seen.

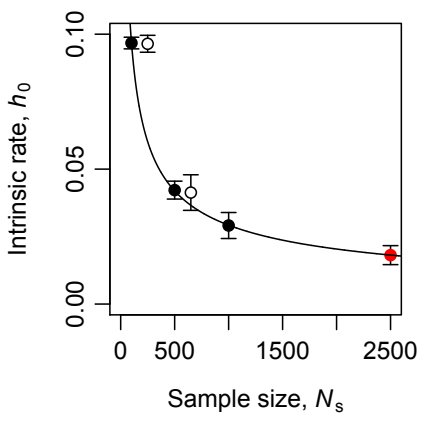

Figure S2.18. The intrinsic rate of evolution h_0 (as a mean of the median intrinsic rates of sets of realizations) as a function of sample size, N_s. Error bars show standard deviations. Black points are based upon the main model realizations, with population size $N_j = 1000$. The black curve shows the best exponential function fit to these three points (see text). The red point is from the supplementary large-population realizations, with $N_j = 2500$ (see Supplemental S2, *Effects of large population size and large sample size*); points for $N_j = 2500$ and $N_s \in \{100, 500, 1000\}$ are not shown because they closely overlap the black points. The white points are from the supplementary small-population realizations, with $N_j = 500$ (see Supplemental S2, *Effects of small population size*); these points are shown horizontally offset since they would otherwise coincide with the black points. Note that since the red and white points were not used in fitting the exponential, they represent independent confirmations of the quality of the fit, and show that the intrinsic rate is essentially independent of the population size, for the range of sizes tested.

The intrinsic rate of evolution

Gingerich (1993) introduced a metric called the "intrinsic rate" of evolution, symbolized h_0, representing the average phenotypic difference between successive generations (standardized by the phenotypic standard deviation). The "random walk" model of Gingerich (1993) assumed an intrinsic rate of 0.1 (with justification from various empirical sources); in our model, on the other hand, the intrinsic rate is an emergent property of the model. We therefore calculated the intrinsic rate of each realization, to assess its emergent value.

Following Gingerich (1993), the intrinsic rate from one generation ("generation 0") to the next ("generation 1") is given by

$$h_0 = \left| \left(\overline{t_1} - \overline{t_0} \right) \Big/ \sigma_{t_0} \right| \qquad \text{(Formula S2.5)},$$

where $\overline{t_0}$ and $\overline{t_1}$ are the mean trait values for generations 0 and 1, respectively, and σ_{t_0} is the standard deviation for the trait in generation 0. The trait here may be neutral or selected, and may entail genetic or phenotypic values; we calculated the intrinsic rate for all four possibilities (a_s, z_s, a_n, z_n). The means and the standard deviation were computed at the beginning of each generation, prior to random mortality (see Supplemental S1, *Process overview*).

For each realization of the model, with 50,000 intergenerational intrinsic rates, we then calculated the overall intrinsic rate. According to Gingerich (1993), this should be the average of the intergenerational intrinsic rates. However, the distribution of these rates was highly skewed because a small minority of generations with phenotypic variance close to (or even equal to) zero led to extremely large (or even infinite) intergenerational intrinsic rates. We chose to discard infinite intrinsic rates for three reasons: they couldn't be handled by our analysis, they were rare (for most realizations), and they would not occur in nature (for quantitative traits). For essentially the same reasons, we also chose to summarize the intrinsic rate across the generations of each realization using the median, rather than the mean, so that occasional outliers would not distort the analysis.

Given a (median) intrinsic rate for each realization, an ANOVA of the intrinsic rate as a function of all of the independent variables (V_E, ω, σ_c, m, T, N_s, G, α, μ, C) showed that the sample size, N_s, explained the vast majority of the variance ($\eta^2 = 98.5$, $P < 0.001$); although some other terms were significant, none explained more than 0.5% of the variance (not shown). An ANOVA containing all two-way interactions (45 additional terms) explained 99.7% of the variance; exactly the same variance was explained by N_s, and no interaction explained more than about 0.1% of variance (not shown).

The observed median intrinsic rate, in other words, was almost completely determined by the size of the sample taken from the population (Fig. S2.18). Furthermore, this relationship was closely fitted by the exponential function (linear regression of log-log transformed values, $t_1 = -103.84$, $P = 0.0061$, adj. $R^2 = 0.9998$; Fig. S2.18, black line)

$$h_0 = 1.06 \times N_s^{-0.520} \qquad \text{(Formula S2.6).}$$

Fitting an exponential function to three data points is, of course, not very impressive. However, this exponential fit almost perfectly predicted the intrinsic rate for a sample size $N_s = 2500$ from the realizations discussed in Supplemental S2, *Effects of large population size and large sample size*; no realizations with $N_s > 1000$ were used to calculate the exponential fit, so this represents a strong independent confirmation of the exponential relationship. Furthermore, the supplementary realizations with $N_j = 2500$, which were not used to produce the exponential fit described above, yielded a fit so close to the first that it could not be shown in Fig. S2.18 because it was visually co-incident (linear regression of log-log transformed values, $t_2 = -99.73$, $P < 0.001$, adj. $R^2 = 0.9997$). Similarly, the observed intrinsic rates from the supplementary realizations with $N_j = 500$, discussed in Supplemental S2, *Effects of small population size*, were also predicted almost perfectly by the exponential fit (Fig. S2.18, white points; note these points have been offset horizontally to allow them to be distinguished).

Although surprising at one level, these findings accord with some theoretical predictions. First, the exponential function approaches zero asymptotically for sufficiently large N_s, expressing the common-sense idea that an infinitely large population under stabilizing selection, fully sampled, would have an intrinsic rate of zero. As the sample size decreases, sampling error is introduced which leads to variance in the estimated population means and standard deviations, and this variance leads to progressively larger estimates for the intergenerational intrinsic rate (keeping in mind the absolute value in Formula S2.5, which is the reason that this sampling error does not average out to zero). For finite population sizes, the largest possible sample size equals the population size, and so the expected true intrinsic rate is non-zero; for $N_j = N_s = 1000$, this value is 0.029, for example, while for $N_j = N_s = 2500$, it is 0.018, and for $N_j = N_s = 500$, it is 0.042 (Fig. S2.18). The population size itself appears to be irrelevant, at least for sizes ≥ 500, as seen from the ability of the data points from $N_j = 1000$ realizations to predict the intrinsic rate for both the $N_j = 500$ and $N_j = 2500$ supplemental realizations; the sample size N_s appears to be the relevant fact that allows prediction of those points, illustrating that sampling error is the mechanism driving the estimated intrinsic rate even when the "sample" is a full population census. The exponential relationship between h_0 and N_s is presumably derivable analytically from known sampling theory, although we will not delve into that here.

The aspect that we find particularly surprising is that all of the other model parameters have such a small effect upon the intrinsic rate; it is interesting that higher mutation rate, for example, does not produce a larger intrinsic rate even though one would expect that it would provide greater genetic variance and therefore greater potential for drift from generation to generation. Perhaps the greater potential for drift is cancelled by the greater genetic variance because the intrinsic rate is standardized by the variance, and so parameters that increase the stochastic fluctuations in the mean also increase the variance and thereby produce the same standardized intrinsic rate. Further theoretical exploration of this observation would be worthwhile, as it may imply that estimates of intrinsic rates should be quite predictable from sample size alone, and that true intrinsic rates should be quite predictable from population size alone.

Because sample size is (according to our model) by far the most important predictor of the estimated intrinsic rate of a population under stabilizing (and also squashed stabilizing) selection, a comparison of empirical intrinsic rates to the predictions of Formula S2.6 would be interesting. The intrinsic rate might, for example, provide a test for whether a population is under some form of stabilizing or squashed stabilizing selection. Conversely, it might also turn out that the intrinsic rate is essentially the same regardless of the selective regime, or regardless even of ecology; and that, too, would be a very interesting result.

These findings provide a new perspective on the results of Gingerich (1993), while underlining the importance of sampling error in the estimation of intrinsic rates. As noted above, Gingerich (1993) estimated that a typical intrinsic rate is 0.1; according to our analysis, this suggests that the typical sample size upon which this estimate was based was roughly 100 individuals. However, even large samples taken over many generations can

produce a substantial overestimate of the true intrinsic rate. Indeed, Fig. S2.18 shows results derived from 50,000 generations per population (and many populations per data point); for samples taken on one or a few populations over only a few generations, sampling error will lead to much larger variance. Separating the variance due to the finite size of the population (which leads to a true non-zero intrinsic rate) from the variance due to sampling (which leads to an overestimation of that true rate, as discussed above) to produce good estimates of true empirical intrinsic rates will require further work on both the empirical and theoretical side of this problem.

Temporal variation in selection

Siepielski et al. (2009) analyzed a dataset of temporally replicated studies of selection in nature, concluding that "selection varies considerably among years, including differences in strength, direction and likely form". Further analysis of a subset of this dataset by Morrissey and Hadfield (2012) led them to conclude that selection is in fact "remarkably consistent in time," although it may still be important in many particular systems and over longer timescales. Although this is ultimately an empirical question, we analyzed the results of our model in a similar manner to these previous studies, to compare temporal variation in selection in our model to that observed in nature, to provide a theoretical perspective on the question.

In our analysis, we followed the general approach of Siepielski et al. (2009) and Morrissey and Hadfield (2012), but we used different statistical methods to better handle deviations from normality and the presence of outliers in the dataset. In particular, rather than using means, we used medians, and rather than standard deviations, we used the median absolute deviation (MAD; see Methods, *Data analysis*).

The typical strength of selection, without regard for its direction, can be assessed using the median of the absolute values of the selection gradients, $|\beta|$ and $|\gamma|$, for each realization. The frequency distributions of these metrics across all realizations of our model (Figs S2.19a–b) showed a roughly negative exponential distribution, in qualitative agreement with the results of Siepielski et al. (their Fig. 2). In particular, relatively weaker selection was more common than stronger selection, for both linear and quadratic selection. However, the empirically observed pattern contains relatively few estimates close to zero (perhaps due to publication bias), and indicates stronger selection overall than was observed for the parameter space explored in our realizations.

Temporal variation in selection can be assessed using the MAD of a temporal series of estimated selection gradients. The distributions of this metric across all realizations of our model, for both linear and quadratic selection (Figs. S2.19c–d), agreed qualitatively with the results of Siepielski et al. (their Fig. 4): relatively small magnitudes of temporal variation in selection were more common than large magnitudes of temporal variation. The distributions roughly followed a negative exponential shape, however, whereas Siepielski et al. observed a somewhat different (more Poisson-like) shape.

Temporal variation specifically in the strength of selection, without regard for its direction, can be measured as the MAD of the absolute values of the selection gradients of a temporal series. The distributions of this metric across all realizations of our model, for linear and quadratic selection (Figs. S2.19e–f), agreed qualitatively with the results of Siepielski et al. (their Fig. 5), with small magnitudes of temporal variation in selection strength being more common than large magnitudes, following a roughly negative exponential distribution.

Siepielski et al. (2009) noted that the strength of selection ($|\beta|$ or $|\gamma|$) tended to be correlated with the magnitude of temporal variation in the strength of selection (the SD of $|\beta|$ or $|\gamma|$), showing that stronger selection tends to also be more temporally variable (their Fig. 6). Our data showed this general relationship also (Figs. S2.20a–b), and many realizations showed a close fit to a linear relationship. For the strength of directional selection, two linear relationships were observed (Fig. S2.20a); realizations on the upper linear relationship used the quantitative genetic architecture, were analyzed on their genetic (not phenotypic) trait values, and involved competition with a narrow fitness function ($\omega = 1.0$), a lower mutation rate ($\mu = 0.00001$), and higher random mortality ($m = 0.5$). The combination of all of these conditions invariably placed a realization on the upper line, representing the highest

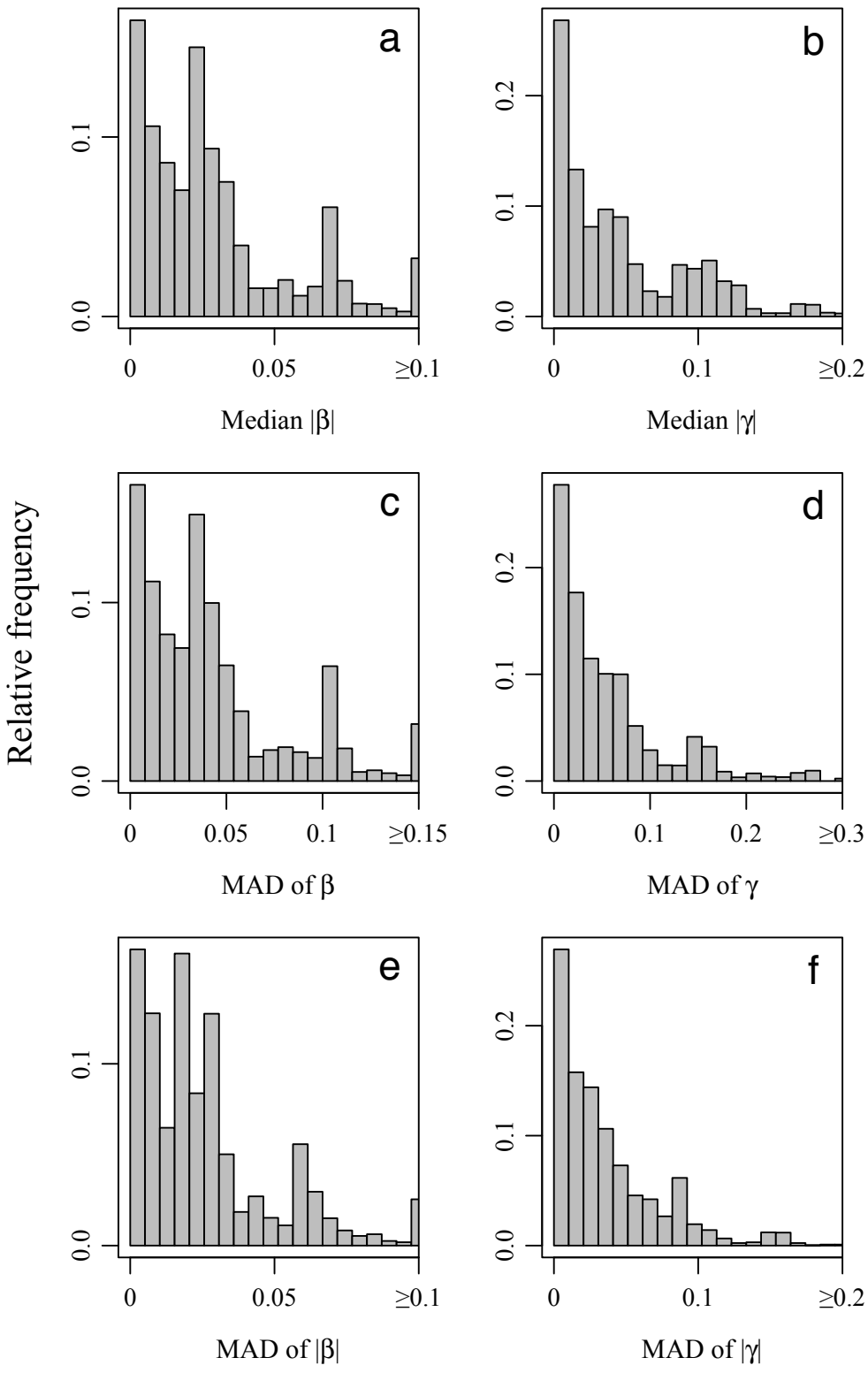

Figure S2.19. Strength and temporal variation of selection. Panels show the frequency distribution of per-realization metrics: (a) median $|\beta|$, (b) median $|\gamma|$, (c) median absolute deviation (MAD; see *Data analysis*) of β, (d) MAD of γ, (e) MAD of $|\beta|$, (f) MAD of $|\gamma|$. In all panels, values greater than the maximum visible bin have been counted in the last visible bin.

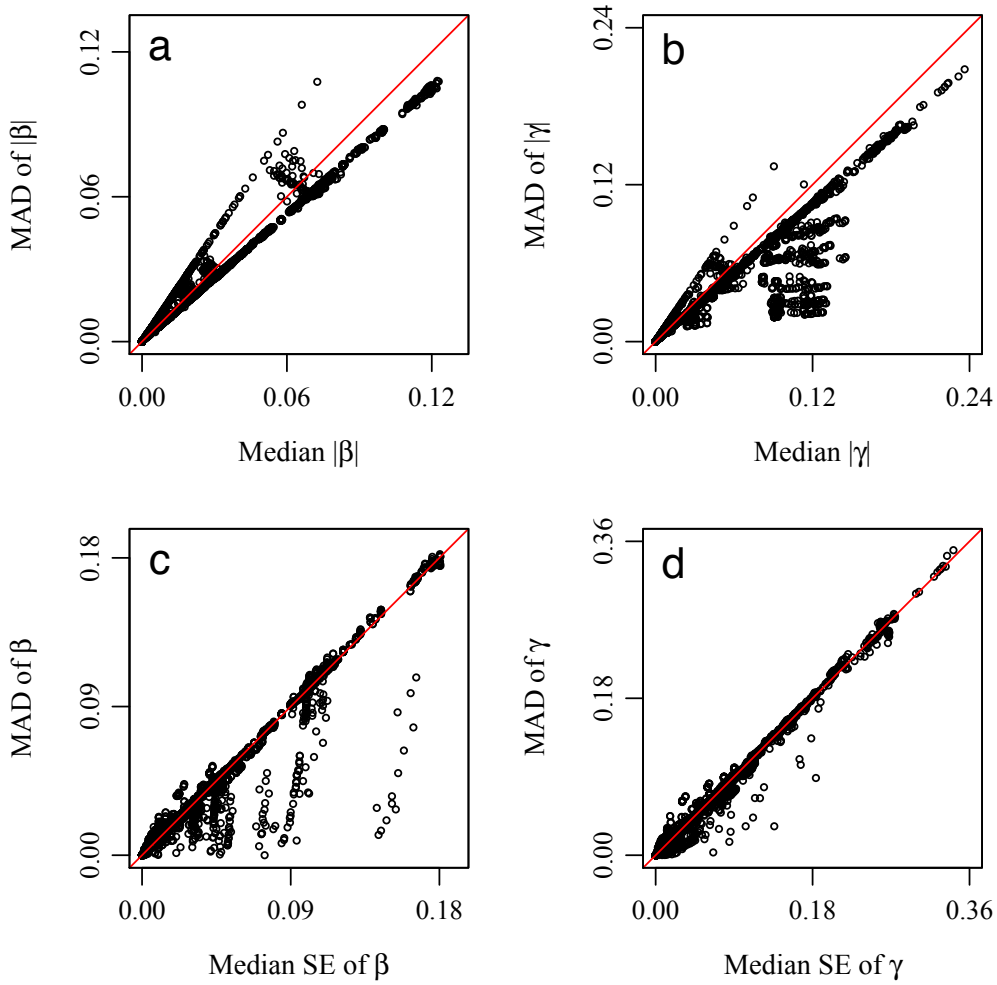

Figure S2.20. Temporal variation in selection. Panels (a) and (b): each point shows, for one realization, the median of the absolute values of the estimates of β (γ) against the median absolute deviation (MAD; see *Data analysis*) of the estimates of β (γ). Panels (c) and (d): each point shows, for one realization, the median standard error of all estimates of β (γ) against the MAD of the estimates of β (γ). Red lines indicate the 1:1 line in all panels. Median standard errors for γ shown in (c) were obtained by doubling the median standard errors from quadratic regression, following Stinchcombe et al. (2008) (see *Data analysis*).

level of temporal variation in selection, relative to the strength of selection, observed for the parameter values used in our realizations. Realizations satisfying some, but not all, of these conditions often fell on the "bridges" between the two lines, while realizations satisfying fewer than a critical subset of these conditions fell on the lower line, representing the lowest level of temporal variation in selection relative to selection strength that we observed.

Two distinct linear relationships were also observed for quadratic selection, but in this case many more realizations did not lie on either of the lines (Fig. S2.20b). Realizations with the quantitative architecture, genetic trait values, a narrower fitness function, lower mutation rate, and higher random mortality again fell exclusively along the upper line; however, for quadratic selection many other realizations also fell on this upper line, indicating that the conditions necessary for high temporal variability relative to selection strength are relaxed compared to the directional selection case discussed above. No single parameter predicted in which region in the plot (the upper line, the lower line, or the lowermost clumps) a realization would

lie. Conversely, the region in which a point was located did not reliably predict any parameter's value, with one exception: all points on the upper line represented realizations with a narrow fitness function.

Morrissey and Hadfield (2012) called many of the results of Siepielski et al. (2009) into question with a re-analysis of their dataset. In particular, Morrissey and Hadfield (2012) asserted that the standard deviation of the selection gradients in most studies is of comparable magnitude to the mean of the standard errors of the selection gradient estimates, and that much of the apparent temporal variation in selection could therefore be accounted for by sampling error, rather than true temporal variation (their Fig. 1). The same analysis performed for our model realizations may be seen in Figs. S2.20c–d. Many realizations fell extremely close to the 1:1 line, and the observed temporal variation in such realizations is indeed likely due to sampling error, as asserted by Morrissey and Hadfield (2012). The nearly exact fit of the points to the 1:1 line was a result of the 50,000 generations of data in each realization, which caused the MAD : median SE ratio due to sampling error to converge towards one. However, even fairly small deviations from the 1:1 line might represent temporal variation in selection greater than (or, curiously, less than) can be accounted for by sampling error; again, the fact that 50,000 generations of estimates are available for each realization makes substantial departure from the 1:1 line due to sampling error alone unlikely. Furthermore, some deviations from the 1:1 were not small; in some cases the MAD is approximately double the median SE.

Further analysis (not shown) supported the idea that even slight departures from the 1:1 line were meaningful here. Low outliers in Fig. S2.20c were generally high outliers in Fig. S2.20a; the same combination of parameters that led to a high MAD absolute β relative to median absolute β also led to a low MAD β relative to the median SE of β. High outliers in Fig. S2.20c, on the other hand, were associated with the absence of competition, with analysis of phenotypic (not genetic) values, and with higher mutation rates, a narrower fitness function, and lower random mortality. These conditions are all the opposite of the conditions that predicted low outliers in Fig. S2.20c, except for a narrower fitness function (which is associated with being an outlier in both directions). Similar patterns hold for Fig. S2.20d (not shown); high outliers are associated with the absence of competition, with analysis of phenotypic values, with higher mutation rate, and with lower random mortality, while lower outliers are associated with the opposite (and again, a narrow fitness function is associated with being an outlier in both directions). In summary, there seem to be specific, predictable conditions for which temporal variation in selection is detected at substantially higher than the rate expected from sampling error alone.

It could be argued that even this is not conclusive evidence for temporal variation in selection in our model; if the phenotypic distribution of the population being sampled is non-normal, for example, that departure from normality would violate the assumptions of the regressions underlying all of this, and that violation could cause systematic biases in the magnitude of the selection gradient estimates versus their associated standard errors. Those systematic biases might be predicted by various parameters, since the parameters would affect the shape of the population's trait distribution, and thus all of the patterns discussed here could conceivably be artifacts. One argument that this is not the case comes from our analysis of temporal autocorrelation (see *Autocorrelation and reproducibility*): if all of the variation in selection observed between generations were driven by sampling error alone, rather than true temporal variation in selection, it is difficult to see how significant positive autocorrelation in selection could be observed to persist for many generations, as we commonly observed. A second argument is logical: with competition, the population often occupies a fitness minimum and experiences disruptive selection (see Discussion, *Patterns of selection with competition* and *Squashed stabilizing selection*). If this fitness landscape were invariant, the population would escape the fitness minimum and evolve toward one of the local fitness peaks; the fact that it does not do so demonstrates that the fitness landscape must be varying temporally, and that fitness landscape variation must be expressed in the actual pattern of selective deaths (since if temporal changes in the fitness landscape were not "enforced"

via selective deaths, the population could again escape the fitness minimum).

Nevertheless, the magnitude of the true temporal variation in selection, beyond sampling error, that we observed was often small. Furthermore, its magnitude relative to the median standard error was largest when its absolute magnitude was small; there were realizations for which the MAD was more than double the mean standard error, for both β and γ, but those realizations lie very close to the lower left corner of Figs. S2.20c–d, and were associated with a large sample size ($N_s = 1000$). Detecting this sort of temporal variation empirically would likely require both large sample sizes and extensive temporal replication.

Most importantly, temporal variation in selection in our model occurred despite a lack of temporal variation in the fitness landscape; in particular, high MAD values were observed without competition, as discussed above, which supports our postulate that populations wander on fixed fitness peaks and thereby experience temporally varying selection (see Introduction). Temporal variation in the underlying fitness landscape must also cause selection to vary temporally in nature, of course, which we do not here model (apart from the frequency-dependent fluctuations of SSS); that variation might be of larger magnitude and thus more readily detectable.

Effects of large population and sample size

An additional 216 realizations were conducted with a larger population size of $N_j = 2500$: 3 genetic architectures × 2 values of C × 3 values of V_E × 2 values of ω × 2 values of σ_c × 3 values of m (with redundancy involving σ_c without competition; see Methods, *Data collection*). These realizations were not used in the analysis presented in the main paper; they are examined here. Only the highest mutational variance for each genetic architecture (i.e. $\mu = 0.001$ and, where applicable, $\alpha = 0.5$) was used, principally because the disk space (760 GB, compressed) and the computation and analysis time (nearly two months) for these realizations prevented further exploration. For the analysis of these supplemental realizations, a full population census ($N_s = 2500$) was used in addition to the subsample sizes of 100, 500, and 1000 used for the main realizations.

Results obtained with a larger population size ($N_j = 2500$) were essentially indistinguishable from the main analysis results ($N_j = 1000$). Comparisons were conducted between all $N_j = 2500$ realizations with sample size $N_s \neq 2500$ and the subset of $N_j = 1000$ realizations with exactly the same parameter values. Each dataset contained 2592 observations, half of which were on the neutral trait, half on the selected trait. Welch's unpaired t-tests were used to compare these datasets, subdivided by neutral/selected trait, against each other, testing for a difference between the means of four metrics: (1) median β, (2) median γ, (3) rate of significance of β, P(β*), and (4) rate of significance of γ, P(γ*). None of these eight tests were significant at a Bonferroni-corrected α level of 0.00625 (not shown). For the selected trait, the difference between median β values would have been significant at $\alpha = 0.05$, but the actual difference was vanishingly small ($N_j = 1000$: mean = -3.3×10^{-5}, $N_j = 2500$: mean = 7.7×10^{-6}, $t_{2566} = -2.68$, $P = 0.0073$). Population size therefore made no appreciable difference to the model dynamics observed, at least for the two population sizes studied (see also *Effects of small population size*).

The other question that might be asked with the larger population size realizations is: did a sample size of $N_s = 2500$ make a difference to the results, as compared to the sample sizes of 100, 500, and 1000 used in the main results? Results from analysis of the $N_j = 2500$ realizations are shown in Figs. S2.21 and S2.22. Large sample size did increase the rate of detection of both linear selection, P(β*) (Fig. S2.21a, Fig. S2.22) and quadratic selection, P(γ*) (Fig. S2.21b, Fig. S2.22), although for P(β*) the median detection rate remained very low. Large sample size also decreased the magnitude of gradient estimates (Fig. S2.22).

These results continue the trend observed for the effects of sample size in the main realizations. As shown in Results, the sample size taken from the population was important to the rate of detection of selection (Figs. 3f, 4f), as postulated, for both linear and quadratic selection, and both without and with competition (Tables S2.1–S2.8). From the perspective of the signal-to-noise ratio (see *Selective deaths and the detection of selection*), subsampling results in a decrease in signal, because it creates the possibility of missing one or more of the selective

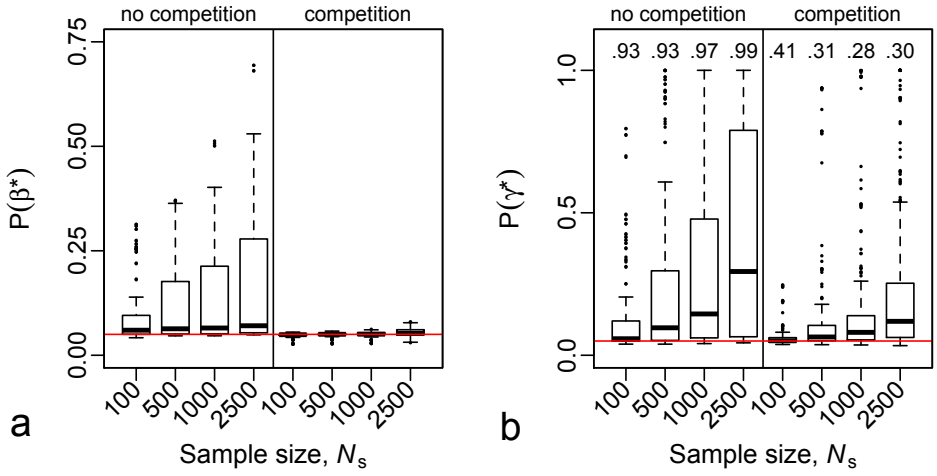

Figure S2.21. Large sample size, $N_s = 2500$, compared to smaller sample sizes, showing effects on: (a) the rate of detection of linear selection, P(β*); and (b) the rate of detection of quadratic selection, P(γ*). These plots may be compared to Figs. 3f and 4f, respectively. Note, however, that the dataset analyzed here is the $N_j = 2500$ supplemental dataset, and thus involves only high mutational variances (see Supplemental S2, *Effects of large population and sample size*); this is the reason for the large differences between the plots here and the corresponding main results. See the captions of Figs. 3 and 4 for presentation details.

deaths that would have been necessary to statistically establish the pattern of selection. However, it is worth noting here that even the full 2500-individual censuses of the large population size realizations presented here did not produce frequent detection of stabilizing selection. The typical (median) realization with a sample size of 2500 still had a rate of detection of quadratic selection of only ~30% without competition and only ~10% with competition (Fig. S2.21b). Directional selection was also found more frequently in this case than with smaller sample sizes (Fig. S2.21a), although if found in addition to quadratic selection, rather than instead of it, this is not necessarily a misleading result. Finally, without competition stabilizing selection was not found frequently in some realizations despite the large sample size, while with competition disruptive selection was still found the majority of the time (Fig. S2.21c). This illustrates that the problems with the empirical detection of stabilizing selection are not simply a consequence of insufficient sample size. Rather, transient dynamics often produce detection of directional or disruptive selection even for a population evolving on a stabilizing fitness function – especially, but not exclusively, when negative frequency-dependent selection is also present.

Effects of small population size

An additional 216 realizations were conducted with a smaller population size of $N_j = 500$: 3 genetic architectures × 2 values of C × 3 values of V_E × 2 values of ω × 2 values of σ_c × 3 values of m (with redundancy involving σ_c without competition; see Methods, *Data collection*). These realizations were not used in the analysis presented in the main paper; they are examined here. Only the highest mutational variance for each genetic architecture (i.e. $\mu = 0.001$ and, where applicable, $\alpha = 0.5$) was used. For the analysis of these supplemental realizations, subsample sizes of 100 and 500 were used (a subsample size of 1000 being impossible since it is larger than the population size).

Results obtained with the smaller population size ($N_j = 500$) were essentially indistinguishable from the main analysis results ($N_j = 1000$), for a given subsample size. Comparisons were conducted between all $N_j = 500$ realizations and the subset of $N_j = 1000$ realizations with exactly the same parameter values. Each dataset contained 1728 observations, half of which were on the neutral trait, half on the selected trait. Welch's unpaired t-tests were used to compare these datasets, subdivided by neutral/selected trait, against each other, testing for a difference between the means of four metrics: (1) median β, (2) median γ, (3) rate of significance of β,

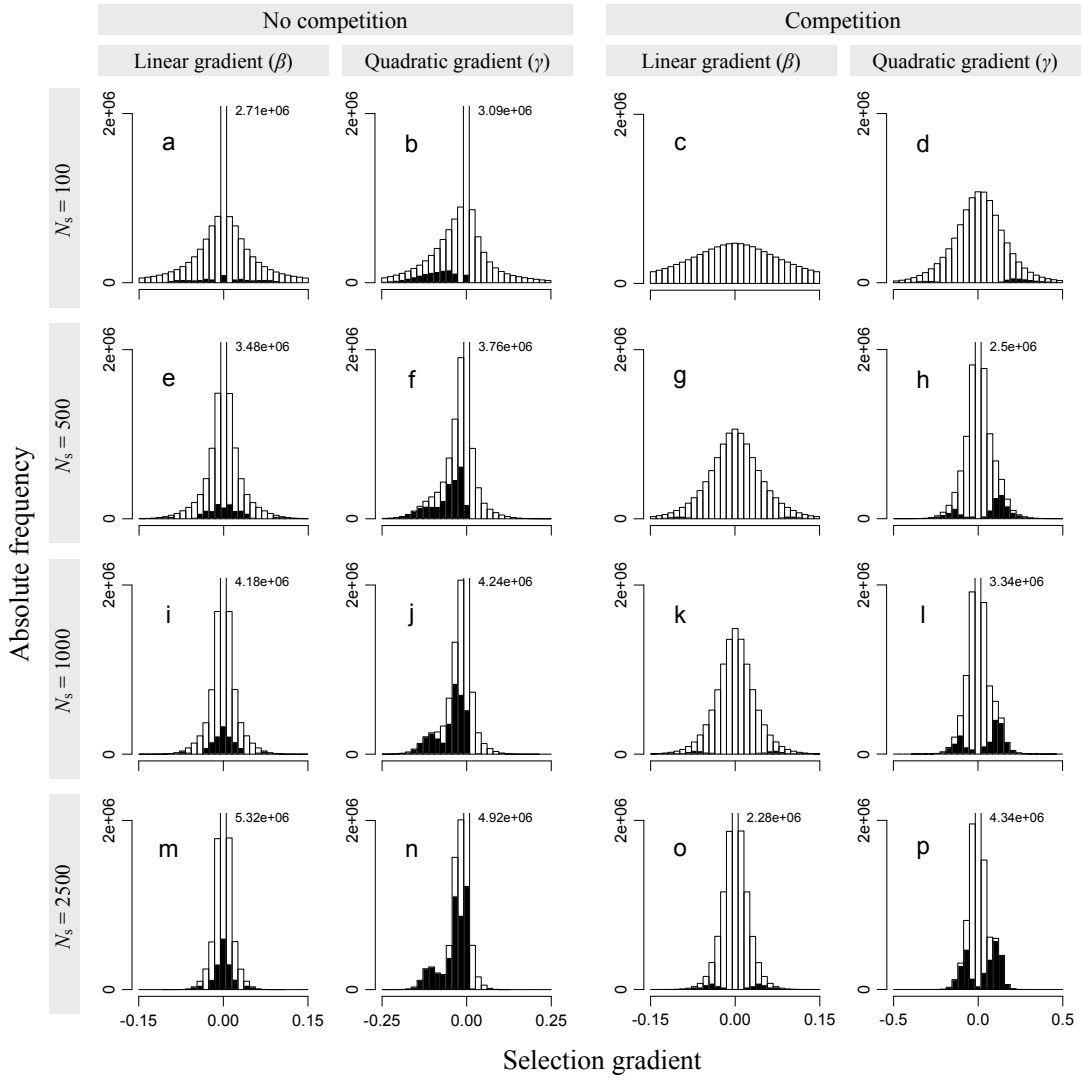

Figure S2.22. Effects of sample size, N_s, on the distributions of estimates of β and γ, for supplemental realizations with large population size ($N_j = 2500$). In all panels, black shading indicates those estimates that are significant ($P < 0.05$). Note x-axis scales are only guaranteed to match within columns. The heights of peaks extending beyond the plot are labeled. Note the supplemental dataset involves only high mutational variances (see Supplemental S2, *Effects of large population and sample size*), causing differences in comparison to Fig. S2.9.

P(β*), and (4) rate of significance of γ, P(γ*). None of these eight tests were significant at a Bonferroni-corrected α level of 0.00625 (not shown). For the selected trait, the difference between median β values would have been significant at $\alpha = 0.05$, but the actual difference was vanishingly small ($N_j = 1000$: mean = -4.5×10^{-5}, $N_j = 500$: mean = 8.7×10^{-6}, $t_{1682} = -2.31$, $P = 0.0212$). Population size therefore made no appreciable difference to the model dynamics observed, at least for the two population sizes studied (see also *Effects of large population and sample size*).

Results from analysis of the $N_j = 500$ realizations are shown in Figs. S2.23 and S2.24, which are parallel to Figs. 3 and 4, although they are not entirely comparable for two reasons. First, Figs. 3 and 4 also include a sample size of 1000, which provides an increased rate of detection of selection. Second, Figs. S2.23 and S2.24, being based on the supplemental realizations described above, only

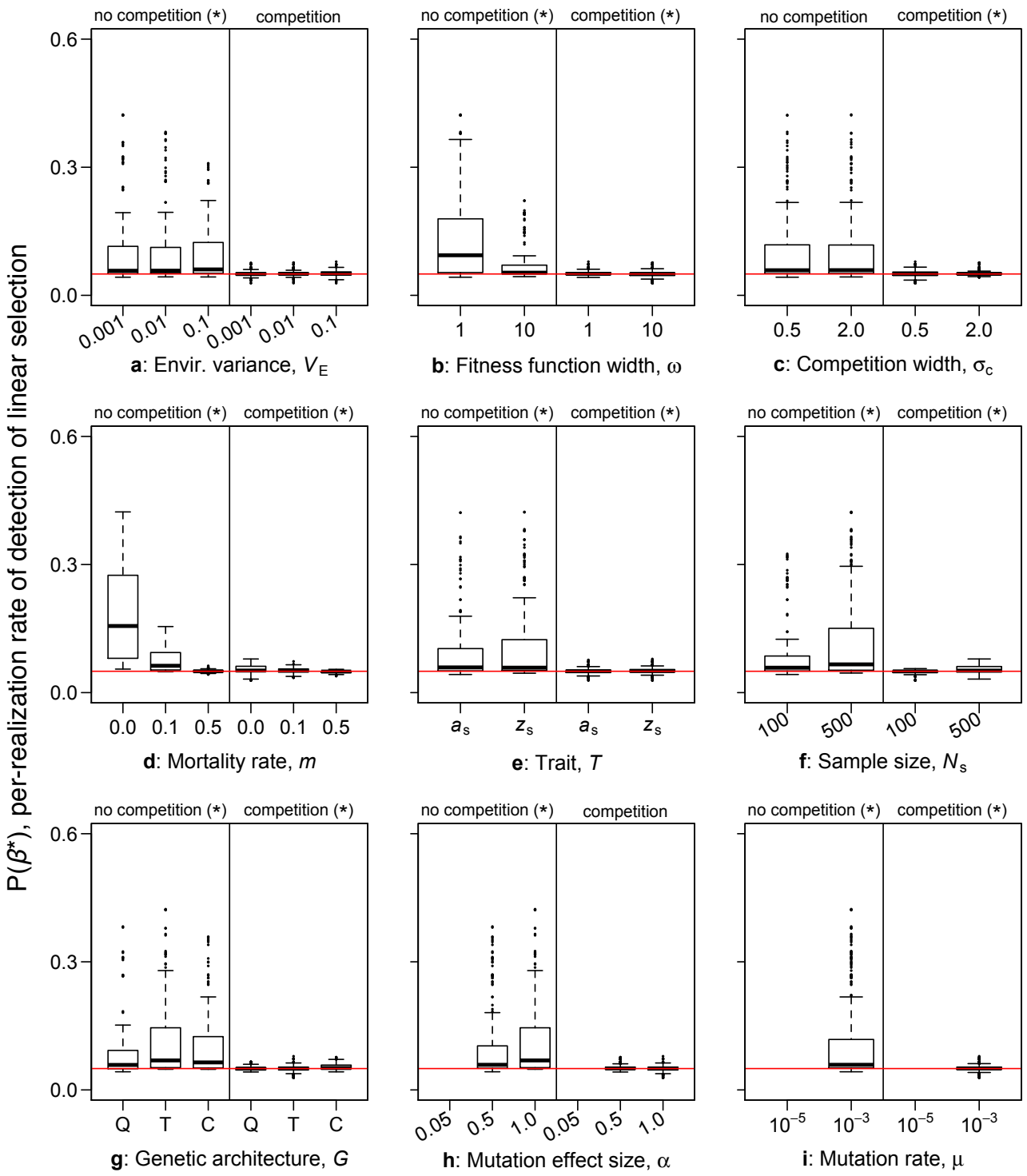

Figure S2.23. Effects of small population size, $N_j = 500$, showing the rate of detection of linear selection, P(β^*), given various parameter values. See Supplemental S2, *Effects of small population size* for discussion, but in short, the results shown here are nearly identical to the results from $N_j = 1000$ realizations with matching parameter values and matching subsample sizes. See the caption of Fig. 3 for presentation details.

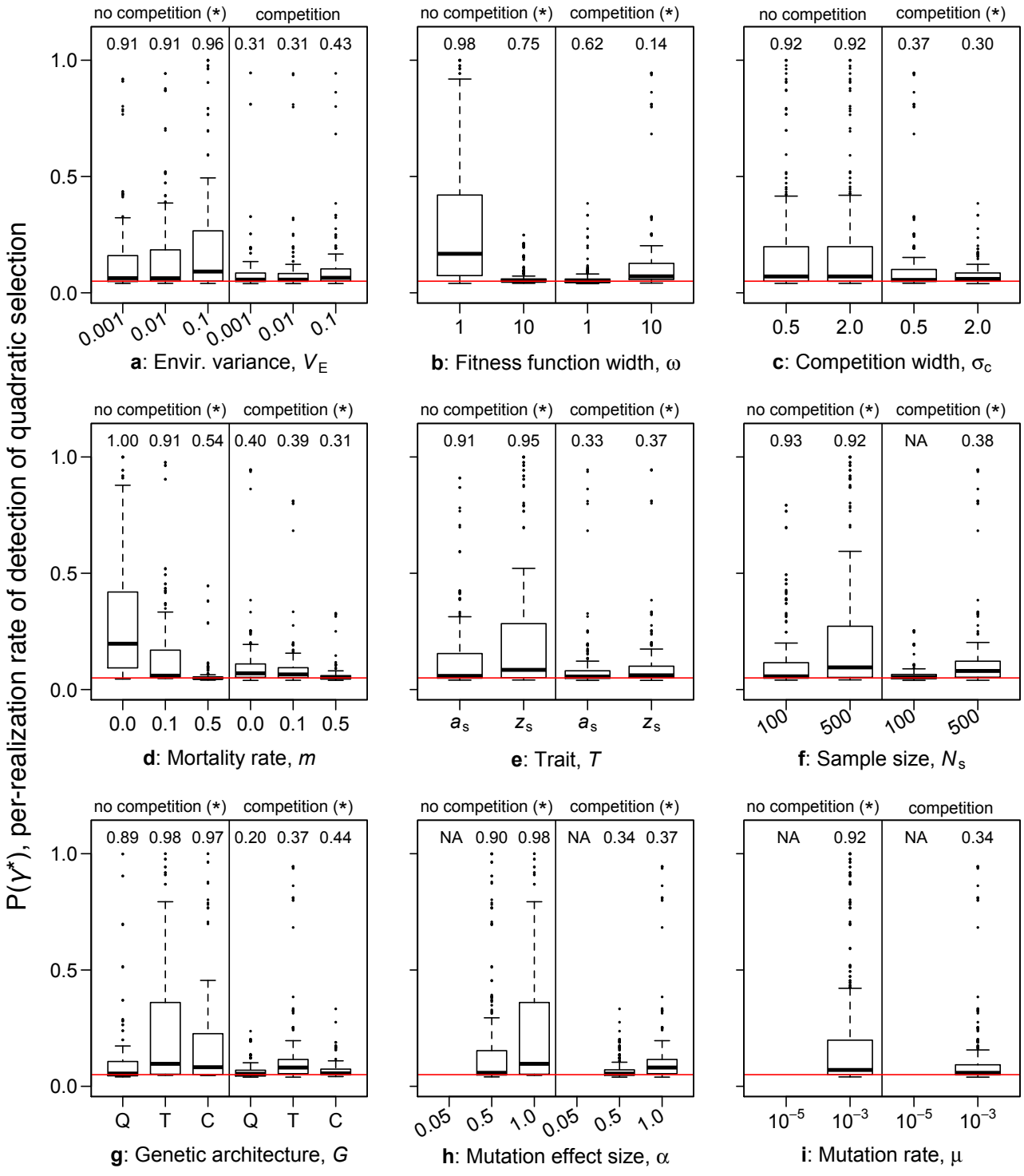

Figure S2.24. Effects of small population size, $N_j = 500$, showing the rate of detection of quadratic selection, P(γ^*), given various parameter values. See Supplemental S2, *Effects of small population size*, for discussion, but in short, the results shown here are nearly identical to the results from $N_j = 1000$ realizations with matching parameter values and matching subsample sizes. See the caption of Fig. 4 for presentation details.

include realizations with a high mutational variance, which changes the model dynamics somewhat. Nevertheless, the overall patterns, and effects of particular parameters, are similar between Figs. S.2.23/S2.24 and Figs. 3/4. Indeed, plots of only the $N_j = 1000$ realizations that match the parameter values used for the $N_j = 500$ realizations, and that exclude the subsample size of 1000, are virtually indistinguishable from Figs. S2.23 and S2.24 (not shown). Small population size thus appears to have essentially no effect, at least down to a population size of 500 individuals. Presumably, at some even smaller population size, drift would begin to play a larger role in the dynamics; but at that point the largest sample size possible would be quite small, too, since it is limited by the overall population size, and so sampling error would obscure such results, and the overall rate of detection of selection would in any case be low.

Estimation of fitness landscape parameters

For each generation of each realization, estimates were calculated for the width of the stabilizing fitness function, ω, and the optimum trait value, θ, from the linear and quadratic selection gradients found by linear regression (β and γ), using the formulas

$$-\left(\omega^2 + \sigma^2\right)^{-1} = \gamma - \beta^2$$

(Formula S2.7; Estes and Arnold 2007)

and

$$\theta = -\gamma^{-1}\beta$$

(Formula S2.8; Phillips and Arnold 1989),

where σ^2 is the phenotypic variance of the population sample. These estimates were then compared against the known values of ω (either 1 or 10) and θ (always 0) to assess the accuracy and precision of these estimation methods.

Three caveats should be noted. First, the β values used in this analysis were from the quadratic regressions (the same regressions from which γ values were obtained), since it would have made little sense to use the linear gradient from one regression and the quadratic gradient from the other; the meaning of β in this section therefore differs from elsewhere in this manuscript. Second, all generations with at least one death (not necessarily a selective death) were used in this analysis; requiring a larger number of deaths, or requiring that β and γ be significant, would have singled out those generations in which selection happened (stochastically) to act more strongly or detectably, and would have thus biased the results. Requiring even one death caused bias, since a lack of deaths is also data, but this was unavoidable since estimates of β and γ are needed for this analysis method. Third, Formula S2.7 is only useful for estimating the strength of stabilizing selection; if selection is instead disruptive, the equation produces an imaginary value (which we did not attempt to interpret). With competition, the detected quadratic selection was predominantly disruptive, making this method inapplicable; we thus present results only from realizations without competition. Even in those realizations, the detected quadratic selection was occasionally disruptive (Fig. 4); for want of a better solution, those generations were discarded from the analysis. This also biased the results, since only generations with a particularly stabilizing pattern of selective deaths were included in the analysis. The consequences of these two biases are discussed below.

This process produced a set of up to 50,000 ω and θ estimates for each realization (one estimate of each parameter for each generation with at least one death and a negative – stabilizing – estimate of γ). Median values were then calculated from these sets, producing per-realization estimates, ω_e and θ_e, of the fitness function width ω and the phenotypic optimum θ respectively. Finally, realizations were grouped by their true fitness function width ω, their sample size N_s, and their random mortality rate m. For each group of realizations, the median and MAD of the per-realization estimates ω_e and θ_e were calculated and plotted (Fig. S2.25). The plotted values thus show the median (with MAD error bars), across a group of realizations, of the per-realization medians, across generations, of the per-generation estimates of ω and θ. Because this process assimilated information from as many as six million generations into each data point, these estimates might be expected to be highly accurate, even if the estimates in individual generations varied widely.

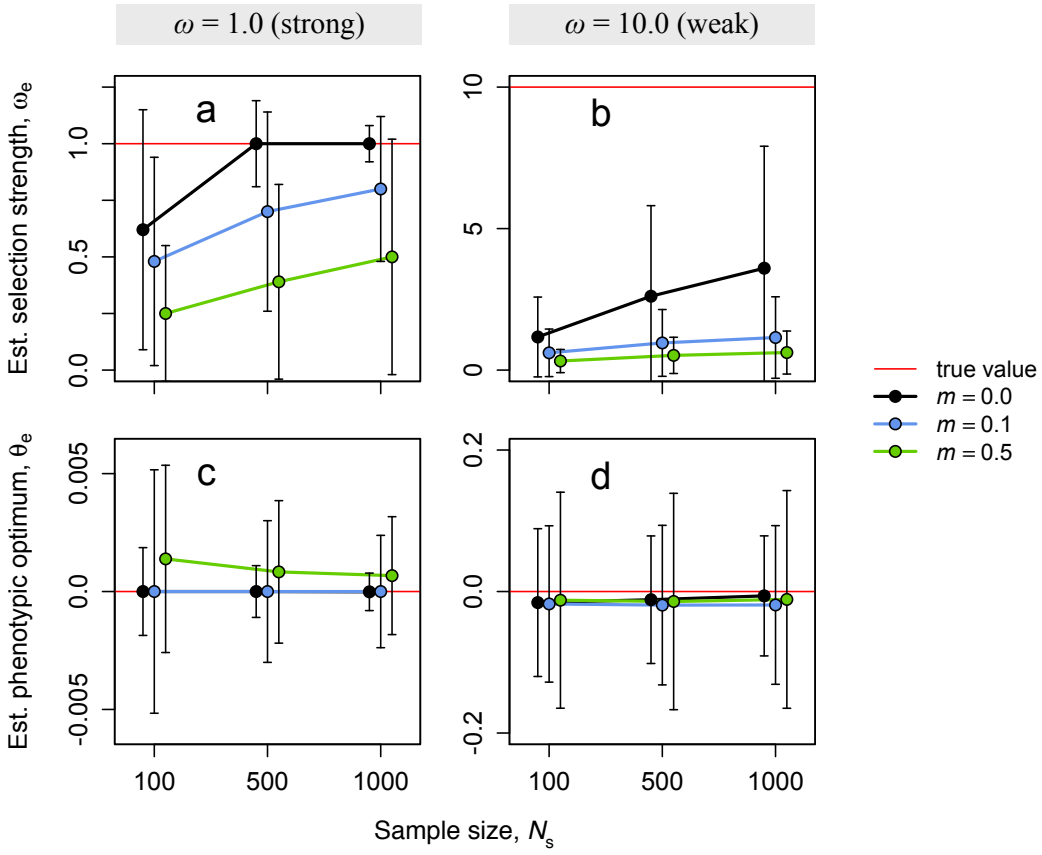

Figure S2.25. Estimates of the shape of the fitness function, derived from selection gradient estimates from realizations without competition. Panels (a) and (b): estimates of the width of the fitness function, ω. Panels (c) and (d): estimates of the position of the phenotypic optimum, θ. Red lines indicate the true values of ω and θ in each panel. Each point shows an estimate (y-axis) derived from a subset of realizations with a particular stabilizing fitness function width ω (columns), a particular sample size N_s (x-axis), and a particular random mortality rate m (colors). For each generation in each realization, an estimate of ω and θ was calculated from that generation's estimated selection gradients (see *Estimation of fitness landscape parameters*). Then for each realization, the median estimates of ω and θ across all generations were calculated. Finally, across each set of realizations represented by one point, the median of these per-realizations estimates was calculated. Error bars show ±MAD (see *Data analysis*). Results shown are for realizations without competition (see text).

Under the most favorable conditions, the estimates were in fact quite accurate. With strong stabilizing selection, no random mortality, and a sample size of 500 or 1000 individuals, the estimates of both ω (Fig. S2.25a) and θ (Fig. S2.25c) were very close to the true value, with little variation among realizations. In other cases, the estimates were often surprisingly inaccurate, as discussed below.

Taking first the estimation of the fitness function width ω, Figs. S2.25a–b show that a general pattern of underestimation of the width of the fitness function (i.e. overestimation of the strength of stabilizing selection) prevailed except, as mentioned above, under the most favorable conditions. Smaller sample size, higher random mortality, and a broader fitness function width (weaker selection) all led to underestimation of the fitness function width. This underestimation might be the result of the biases mentioned above: analyzing only generations in which at least one death occurred, and only generations that produced an estimate that selection was stabilizing rather than disruptive. By discarding information from generations that would suggest that stabilizing selection was weak or even non-existent, these criteria doubtless produced a bias towards overestimation of the selection strength. However, we note that empirical studies would

likely have to follow the same methodology, since the generations that we excluded cannot be used within this analysis framework. The same biases might therefore be expected in empirical studies that estimate the strength of stabilizing selection using this method.

It is notable that the less information was actually available (smaller sample size, higher random mortality), the stronger selection appeared to be. The same trend has been observed in empirical studies of selection (Kingsolver et al. 2001; Hereford et al. 2004), and we observed the same trend in our β and γ estimates also (Figs. S2.7, S2.9; *Effects of parameters on the selection gradient distribution*). The overestimation of the strength of stabilizing selection observed here may therefore not be entirely due to the biases discussed above; they may also be the result of biases introduced earlier in the analytical chain.

The across-realization-group median estimates of the position of the phenotypic optimum θ were generally quite accurate (Figs. S2.25c–d). The variation among realizations within each group, as shown by the MAD error bars, was fairly wide when stabilizing selection was weak, however. Under weak stabilizing selection, few selective deaths occurred in any one generation, and so there was very little information from which to determine the position of the optimum. In many generations, the estimated position of the optimum likely depended mostly or entirely on the pattern of random mortality, not on selective deaths. Since random mortality is, by definition, uncorrelated with phenotype, this would lead, on average, to an estimation that the phenotypic optimum is equal to the mean phenotype of the population – even if the population has, in fact, wandered far from the true optimum. For this reason, a larger sample size has essentially no effect on the variation among realizations for this estimate when random mortality is high (Fig. S2.25d). Low random mortality improves the signal-to-noise ratio and generates estimates closer to the true optimum; on the other hand, most generations are rejected because they contained no deaths at all, so there is little data to go on, and the variation among realizations is still fairly large (but does decrease with larger sample size).

It is worth emphasizing again that every per-realization estimate was generated from 50,000 generations of data. The poor performance of the estimates shown here is thus sobering to contemplate; estimates based on one or a few generations will be less accurate still.

Although the analysis above presents only results from realizations without competition, a little can be said about the case in which negative frequency-dependent selection has been added to produce squashed stabilizing selection. Because of its flattened or dimpled peak, squashed stabilizing selection is often detected as disruptive selection, and is therefore not tractable using Formula S2.7, as discussed above. When stabilizing selection is detected, it is expected to appear weaker than the true strength of the underlying stabilizing fitness function, because the disruptive selection due to negative frequency-dependence opposes stabilization. In such circumstances, the estimate produced by Formula S2.7 will reveal the shape of the composite fitness function, not the underlying stabilizing fitness function alone. Situations in which evolutionary stasis appears to prevail over many generations, and yet the selection detected appears to be either very weak or disruptive, thus fit the expected empirical signature of squashed stabilizing selection. Better methods may be needed to determine the true shapes of fitness functions in the wild (see Discussion, *Comparisons to selection estimates from natural populations*).

Literature Cited

Estes, S., and S. J. Arnold. 2007. Resolving the paradox of stasis: Models with stabilizing selection explain evolutionary divergence on all timescales. Am. Nat. 169:227–244.

Falconer, D. S. 1989. Introduction to Quantitative Genetics. John Wiley & Sons, Inc., New York, NY.

Gingerich, P. D. 1993. Quantification and comparison of evolutionary rates. Am. J. Sci. 293A:453–478.

Hansen, T. F., and D. Houle. 2008. Measuring and comparing evolvability and constraint in multivariate characters. J. Evol. Biol. 21:1201–1219.

Hansen, T. F., C. Pelabon, and D. Houle. 2011. Heritability is not evolvability. Evol. Biol. 38:258–277.

Hereford, J., T. F. Hansen, and D. Houle. 2004. Comparing strengths of directional selection: How strong is strong? Evolution 58:2133–2143.

Houle, D. 1992. Comparing evolvability and variability of quantitative traits. Genetics 130:195–204.

Janzen, F. J., and H. S. Stern. 1998. Logistic regression for empirical studies of multivariate selection. Evolution 52:1564–1571.

Kingsolver, J. G., H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. E. Hill, A. Hoang, P. Gibert, and P. Beerli. 2001. The strength of phenotypic selection in natural populations. Am. Nat. 157:245–261.

Levine, T. R., and C. R. Hullett. 2002. Eta squared, partial eta squared, and misreporting of effect size in communication research. Hum. Comm. Res. 28:612–625.

Morrissey, M. B., and J. D. Hadfield. 2012. Directional selection in temporally replicated studies is remarkably consistent. Evolution 66:435–442.

Mousseau, T. A., and D. A. Roff. 1987. Natural selection and the heritability of fitness components. Heredity 59:181–197.

Siepielski, A. M., J. D. DiBattista, and S. M. Carlson. 2009. It's about time: The temporal dynamics of phenotypic selection in the wild. Ecol. Lett. 12:1261–1276.

Slatkin, M. 1979. Frequency- and density-dependent selection on a quantitative character. Genetics 93:755–771.

Stinchcombe, J. R., A. F. Agrawal, P. A. Hohenlohe, S. J. Arnold, and M. W. Blows. 2008. Estimating nonlinear selection gradients using quadratic regression coefficients: Double or nothing? Evolution 62:2435–2440.