Supplementary material

Supplementary Methods: Macro-environmental effects

In the previous setting, the parameter $\sigma^2_P = \sigma^2_E + \sigma^2_A$ conditions both the phenotypic distribution (the expected phenotypic variance is σ^2_P) and the residual variance of the mean of the population (expected sampling variance σ^2_P/N). However, in some experimental settings, the mean phenotype is expected to vary more than by sampling effects only. For instance, the quality of the environment (food quality, temperature...) can be variable even in controlled lab experiments, changing randomly between generations. Not accounting for such effects may bias the estimate of σ^2_E upwards, and provide an imprecise picture of the variance components.

It is possible to introduce an additional layer of residual variance, by splitting the environmental effects into a micro environmental variance (σ^2_E, as defined previously), and a macro-environmental variance, σ^2_{me}, quantifying the amplitude of generation-specific shifts in the mean phenotype. In this setting, the phenotypic values at generation t are normally distributed around a phenotypic mean m_t, which is itself a random number drawn into a normal distribution of mean μ_t, the expected mean given the genetic architecture model, and of variance σ^2_{me}. Equation (16) thus becomes:

$$\mathrm{P}(Y_t|\mu_t, m_t, \sigma^2_{P_t}, \sigma^2_{me}) = \phi(m_t|\mu_t, \sigma^2_{me}) \prod_{i=1}^{N_t} \phi(y_{it}|m_t, \sigma^2_{P_t}). \tag{22}$$

It is unnecessary, as well as technically challenging, to estimate m_t independently for all generations by maximum likelihood. Indeed, generation-specific means m_t represent nuisance parameters that have no particular interest by themselves, but should be included in the model to avoid biasing other parameters of interest. It is thus possible to consider them as random effects, and to eliminate them from the vector of parameters $\boldsymbol{\Theta}$. The joint distribution of y_{it} and m_t is multivariate normal, hence the marginal distribution of y_{it} multivariate normal, so that

$$\mathrm{P}(Y_t|\mu_t, \sigma^2_{P_t}, \sigma^2_{me}) = \frac{1}{(2\pi)^{N/2}\det(\boldsymbol{V})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}_t - \mu_t)^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y}_t - \mu_t)\right), \tag{23}$$

where $\boldsymbol{y}_t$ is the vector of phenotypic observations at generation t, $^{\mathrm{T}}$ denotes the transposition operation, and $\boldsymbol{V}$ is the $N \times N$ variance-covariance matrix of phenotypic observations:

$$\boldsymbol{V} = \begin{bmatrix} \sigma^2_P + \sigma^2_{me} & \sigma^2_{me} & \cdots & \sigma^2_{me} \\ \sigma^2_{me} & \sigma^2_P + \sigma^2_{me} & \cdots & \sigma^2_{me} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{me} & \sigma^2_{me} & \cdots & \sigma^2_P + \sigma^2_{me} \end{bmatrix}.$$

Given that $\det(\boldsymbol{V}) = (\sigma^2_P)^N + N\sigma^2_{me}(\sigma^2_P)^{N-1}$ and

$$\boldsymbol{V}^{-1} = \frac{1}{\sigma^2_P(\sigma^2_P + N\sigma^2_{me})} \begin{bmatrix} \sigma^2_P + (N-1)\sigma^2_{me} & -\sigma^2_{me} & \cdots & -\sigma^2_{me} \\ -\sigma^2_{me} & \sigma^2_P + (N-1)\sigma^2_{me} & \cdots & -\sigma^2_{me} \\ \vdots & \vdots & \ddots & \vdots \\ -\sigma^2_{me} & -\sigma^2_{me} & \cdots & \sigma^2_P + (N-1)\sigma^2_{me} \end{bmatrix},$$

equation (23) can be rewritten as:

$$\mathrm{P}(Y_t|\mu_t,\sigma^2_{P_t},\sigma^2_{me}) = \frac{1}{\sqrt{(2\pi)^{N_t}(\sigma^2_{P_t})^{N_t-1}(\sigma^2_{P_t}+N_t\sigma^2_{me})}}\cdot$$
$$\exp\left(-\frac{1}{2}\frac{[\sigma^2_{P_t}+(N_t-1)\sigma^2_{me}]\sum_{i=1}^{N_t}(y_{it}-\mu_t)^2-\sigma^2_{me}\sum_{i=1}^{N_t}\sum_{i\neq i}(y_{it}-\mu_t)(y_{jt}-\mu_t)}{\sigma^2_{P_t}(\sigma^2_{P_t}+N_t\sigma^2_{me})}\right). \quad (24)$$

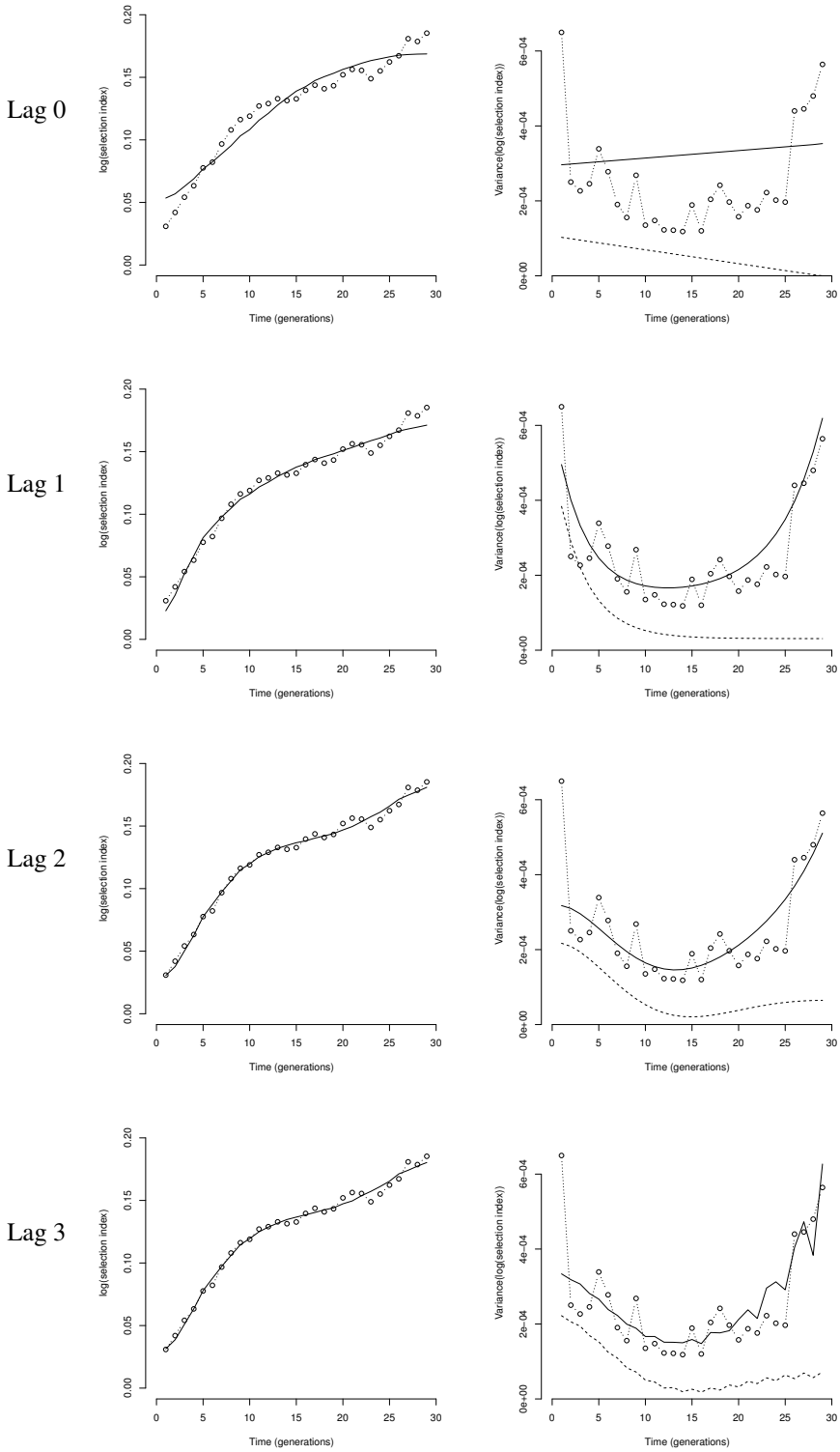
This probability can be expressed in terms of observed mean and variance:

$$\mathrm{P}(\bar{y}_t, s^2_{y_t}|\mu_t,\sigma^2_{P_t},\sigma^2_{me};N_t) =$$
$$\frac{1}{\sqrt{(2\pi)^N(\sigma^2_P)^{N-1}(\sigma^2_P+N\sigma^2_{me})}}\exp\left(-\frac{N_t}{2\sigma^2_{P_t}}\cdot\frac{\sigma^2_{P_t}[s^2_{y_t}+(\bar{y}_t-\mu_t)^2]+N_ts^2_{y_t}\sigma^2_{me}}{\sigma^2_{P_t}+N_t\sigma^2_{me}}\right). \quad (25)$$

As expected, replacing $\sigma^2_{me}=0$ in equation (25) gives equation (17).

Supplementary Figure 1: phenomenological models, 'up 1' line

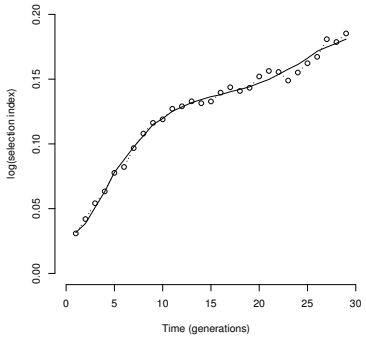
Illustration of the fit of phenomenological models of increasing complexity on the first "up" line. The broken line corresponds to the predicted additive genetic variance.

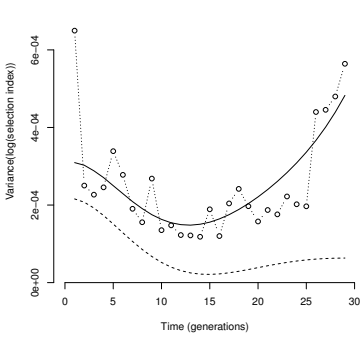


Supplementary Figure 2: Fitting phenomenological models (lag 2) on individual lines.

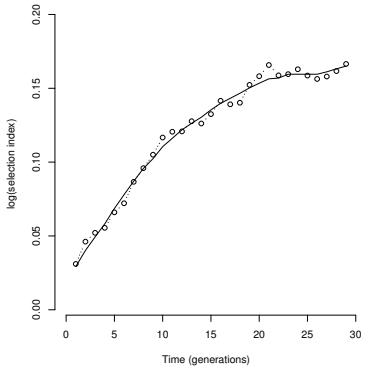
The broken line corresponds to the predicted additive genetic variance.

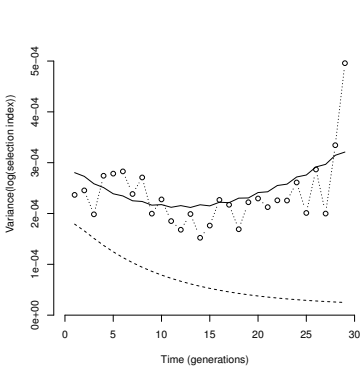
up 1



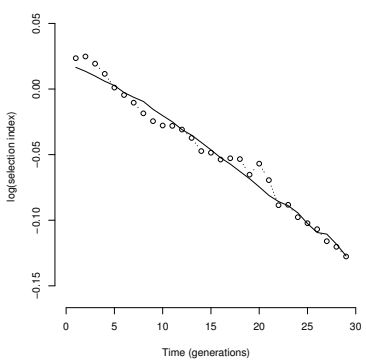


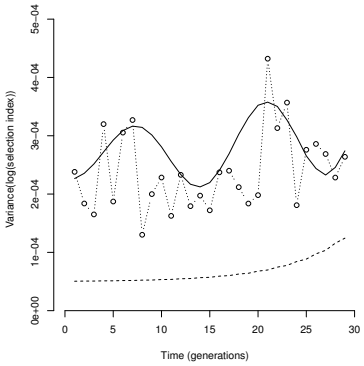
up 2



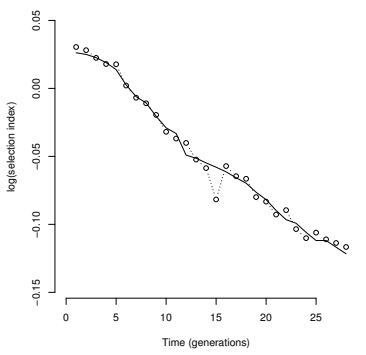


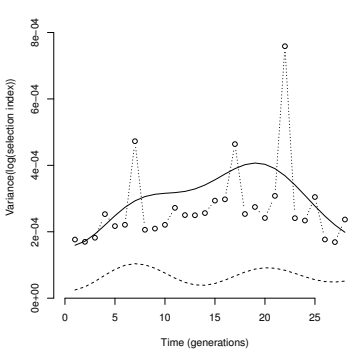
down 1





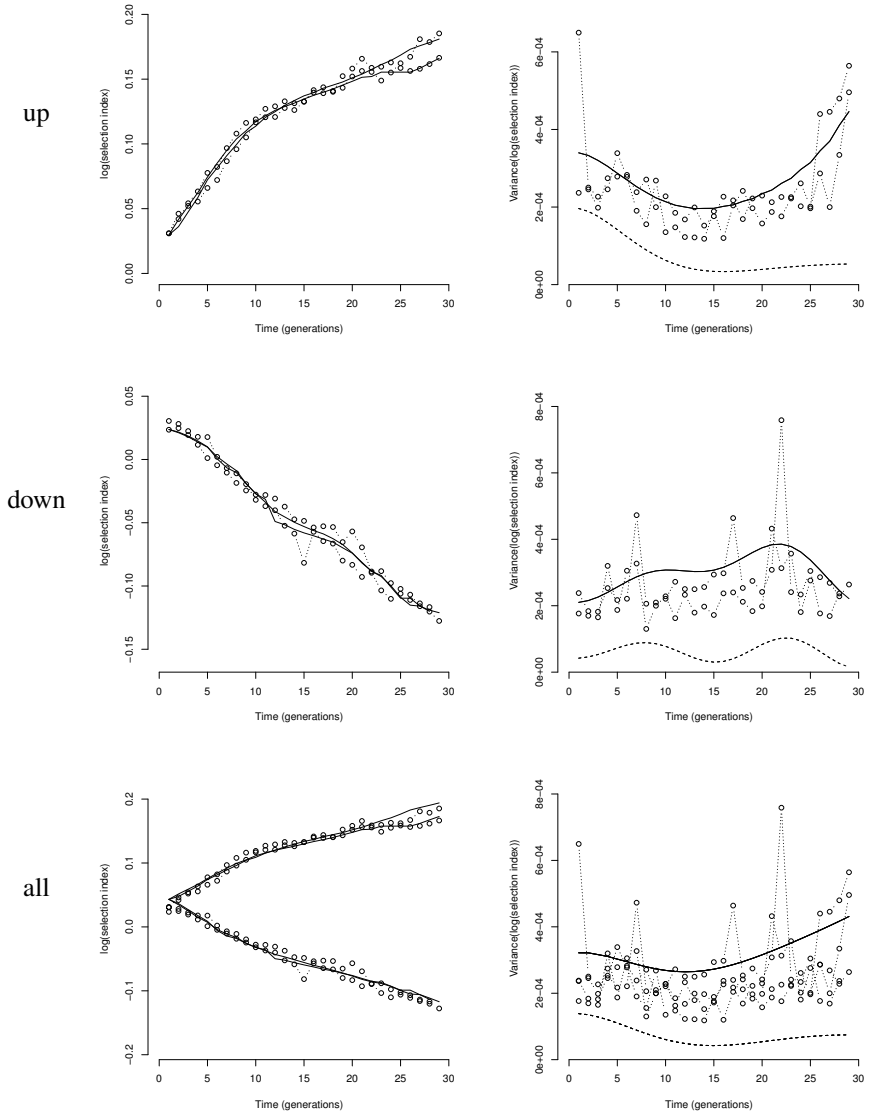
down 2





Supplementary Figure 3: Fitting phenomenological models (lag 2) on combined lines.

The broken line corresponds to the predicted additive genetic variance.



Supplementary Results 1

Precision of the estimates in the constant-variance model (equation (6)) for three population sizes ($N =$ 20, 100, and 1000) and three lengths for the times series ($G = 5$, 10 and 50 generations). 100 simulations were run according to the 'stochastic simulation' procedure described in the Methods section, and for each parameter are indicated the relative bias of the estimate (difference between the mean estimate and the true value), and the standard error (standard deviation of the estimate). Simulated parameter values are $\mu_0 = 0.04$, $\sigma_A^2 = 0.0001$, and $\sigma_E^2 = 0.0004$, selection strength is $\beta = 50$.

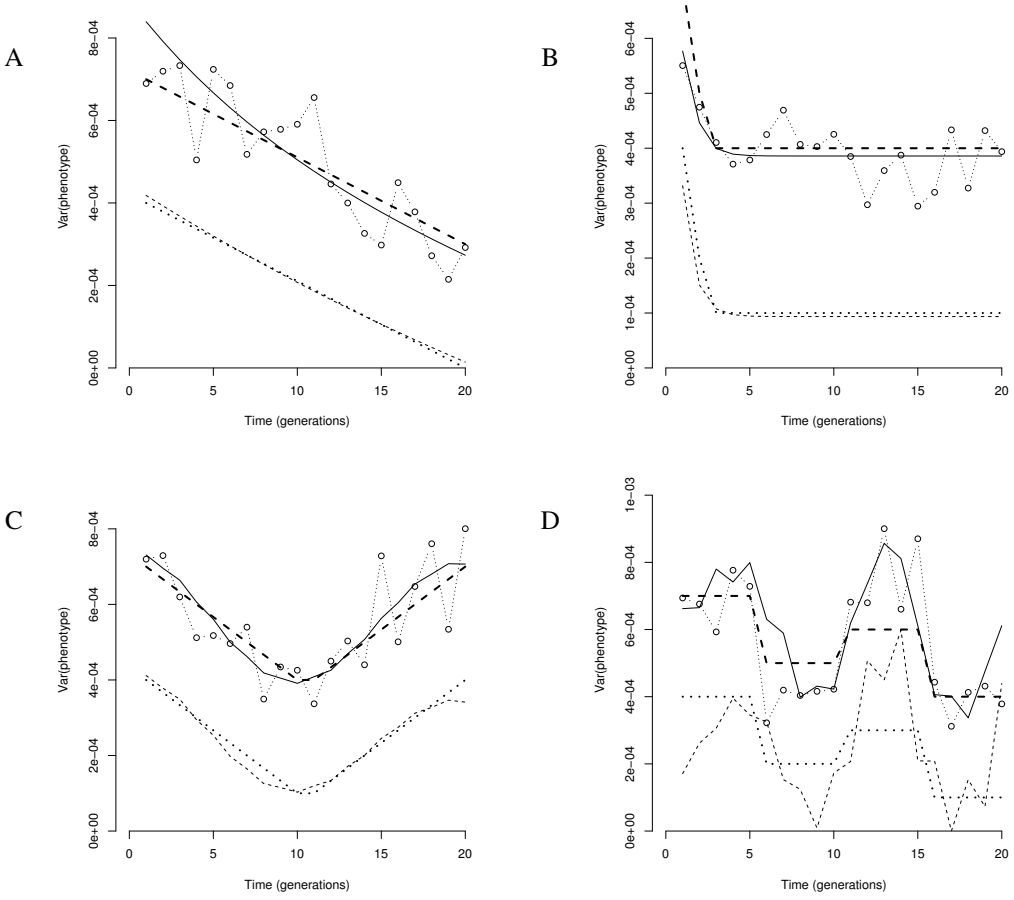
$\mu_0 = 0.04$		Bias: $(\bar{\hat{\mu_0}} - \mu_0)/\sigma_P$	Precision: sd $(\hat{\mu_0})/\sigma_P$
G=5	N=20	0.025	0.159
	N=100	0.006	0.074
	N=1000	- 0.001	0.023
G=10	N=20	- 0.006	0.120
	N=100	0.000	0.059
	N=1000	0.001	0.017
G=50	N=20	0.001	0.056
	N=100	0.002	0.028
	N=1000	- 0.001	0.009

$\sigma_A^2 = 0.0001$		Bias: $(\bar{\hat{\sigma_A^2}} - \sigma_A^2)/\sigma_A^2$	Precision: sd $(\hat{\sigma_A^2})/\sigma_A^2$
G=5	N=20	- 0.001	0.003
	N=100	-0.000	0.001
	N=1000	0.000	0.000
G=10	N=20	0.000	0.001
	N=100	- 0.000	0.001
	N=1000	- 0.000	0.001
G=50	N=20	0.000	0.001
	N=100	- 0.000	0.000
	N=1000	0.000	0.000

$\sigma_E^2 = 0.0004$		Bias: $(\bar{\hat{\sigma_E^2}} - \sigma_E^2)/\sigma_E^2$	Precision: sd $(\hat{\sigma_E^2})/\sigma_E^2$
G=5	N=20	0.001	0.003
	N=100	- 0.000	0.001
	N=1000	- 0.000	0.000
G=10	N=20	0.001	0.003
	N=100	- 0.000	0.001
	N=1000	- 0.000	0.000
G=50	N=20	0.001	0.001
	N=100	- 0.000	0.000
	N=1000	0.000	0.000

Supplementary Results 2

Flexibility of the phenomenological models. The autoregressive-like design of the phenomenological models aims at providing a simple but efficient way to catch the dynamics of genetic and environmental variances during the selection response, without *a priori* constrains on what the variance dynamics are supposed to be. In order to illustrate the flexibility of this framework and to test its limits, some totally arbitrary genetic variance dynamics were simulated in a population of size $N = 100$ under constant directional selection ($\beta = 50$) for 20 generations. The environmental variance is constant ($\sigma^2_E = 0.0004$), while σ^2_A varies as illustrated in the figures. A: linear decay of the variance, B: fast decay in the first generations, followed by an equilibrium (similar to what would be observed with a strong Bulmer effect). C: Arbitrary down and up dynamics, D: complex arbitrary shifts in the additive variance. The figures present the phenotypic variance observed in the simulation (dots), the theoretical phenotypic variance (bold dashed) and the predicted phenotypic variance from the model (plain thin line). The theoretical additive variance (bold dotted line) and the predicted additive variance (thin dashed line) are also represented. Models fit on cases A and B are simple models (only lags 0 and 1), while cases C and D required to involve lags 2 and 3 as well. In D, the phenomenological model manages to catch the general pattern by predicting cyclical changes, but some details of the dynamics are not precisely handled.



Supplementary Results 3

Accuracy of the model selection procedure. Two genetic architectures were considered: (i) a constant-variance model ($\mu_0 = 0.04$, $\sigma_A^2 = 0.0001$, $\sigma_E^2 = 0.0004$) (equation (6)), and (ii) a negative directional epistasis model (same parameter values, plus $\varepsilon = -1$) (equation (12)). Times series of different lengths (G=10 or G=30) were generated from these two models, either only an "up" selected line (T=1, with $\beta = 50$), either two selected lines (T=2, $\beta = 50$ and $\beta = -50$). Sampling stochasticity was then simulated according to the 'stochastic simulation' procedure detailed in the Methods section, with two sample sizes (N=20 or N=100). Six models were then fit to each simulated data sets: the constant-variance model (equation (6)), the drift model (7), the mutation model (including drift) (8), directional epistasis model (12), joint effect selection model (without drift) (14), and a canalization model with both genetic and environmental canalization (equation (13)). Two model selection criteria were considered, AIC and BIC. AIC aims at identifying a good model among a set of candidates, and is therefore adequate when exploring a real data set, while BIC is good at picking the right model among a set of simulated results (and is thus expected to perform optimally in this particular test). Each table presents for two situations (top: constant-variance, bottom: epistasis; 'true' models are indicated in bold) the mean AIC and BIC scores obtained over 100 simulations relative to the best average AIC/BIC (i.e. the best model has a score of 0), as well as the percentage of cases in which the model had the best AIC/BIC score among the six tested.

AIC			**Constant**	Drift	Mutation	Epistasis	Selection	Canalization
T=1	G=10	N=20	0.0 (87%)	1.7 (2%)	3.7 (0%)	3.2 (1%)	4.0 (0%)	2.7 (10%)
		N=100	0.0 (86%)	1.7 (5%)	3.7 (0%)	2.8 (2%)	4.0 (0%)	3.0 (7%)
	G=30	N=20	0.0 (86%)	1.6 (6%)	3.5 (0%)	3.0 (1%)	3.6 (0%)	2.9 (7%)
		N=100	0.0 (85%)	1.8 (3%)	3.7 (0%)	2.7 (3%)	3.9 (0%)	3.1 (9%)
T=2	G=10	N=20	0.0 (79%)	1.8 (2%)	3.8 (0%)	2.3 (8%)	3.9 (0%)	2.6 (11%)
		N=100	0.0 (76%)	1.9 (0%)	3.9 (0%)	1.5 (14%)	4.0 (0%)	2.4 (10%)
	G=30	N=20	0.0 (73%)	1.8 (4%)	3.8 (0%)	1.8 (7%)	2.7 (3%)	1.9 (13%)
		N=100	0.2 (41%)	2.7 (0%)	4.6 (0%)	0.0 (38%)	3.1 (1%)	1.2 (20%)
			Constant	Drift	Mutation	**Epistasis**	Selection	Canalization
T=1	G=10	N=20	0.0 (87%)	1.9 (0%)	3.9 (0%)	2.9 (2%)	4.0 (0%)	2.8 (11%)
		N=100	0.0 (79%)	1.9 (2%)	3.9 (0%)	2.2 (9%)	4.0 (0%)	3.1 (10%)
	G=30	N=20	4.1 (24%)	6.2 (0%)	6.6 (2%)	0.0 (16%)	2.4 (5%)	0.3 (53%)
		N=100	38.4 (0%)	40.5 (0%)	32.5 (8%)	0.0 (35%)	9.0 (6%)	4.9 (51%)
T=2	G=10	N=20	0.0 (81%)	1.8 (3%)	3.8 (0%)	2.4 (9%)	4.0 (0%)	2.9 (7%)
		N=100	1.0 (44%)	2.9 (0%)	4.9 (0%)	0.0 (44%)	4.9 (0%)	3.5 (12%)
	G=30	N=20	109.3 (0%)	111.4 (0%)	113.1 (0%)	0.0 (98%)	14.8 (2%)	107.4 (0%)
		N=100	579.2 (0%)	581.7 (0%)	582.9 (0%)	0.0(100%)	70.8 (0%)	561.7 (0%)

BIC			**Constant**	Drift	Mutation	Epistasis	Selection	Canalization
T=1	G=10	N=20	0.0 (95%)	2.8 (2%)	5.9 (0%)	5.4 (0%)	6.2 (0%)	4.9 (3%)
		N=100	0.0 (94%)	2.8 (2%)	5.9 (0%)	4.9 (1%)	6.2 (0%)	5.2 (3%)
	G=30	N=20	0.0 (96%)	3.7 (3%)	7.8 (0%)	7.2 (0%)	7.9 (0%)	7.2 (1%)
		N=100	0.0 (97%)	3.9 (2%)	8.0 (0%)	7.0 (0%)	8.2 (0%)	7.3 (1%)
T=2	G=10	N=20	0.0 (97%)	3.6 (0%)	7.4 (0%)	5.9 (2%)	7.5 (0%)	6.2 (1%)
		N=100	0.0 (93%)	3.7 (0%)	7.5 (0%)	5.1 (2%)	7.6 (0%)	6.0 (5%)
	G=30	N=20	0.0 (98%)	4.6 (0%)	9.4 (0%)	7.4 (0%)	8.3 (0%)	7.5 (2%)
		N=100	0.0 (96%)	5.3 (0%)	9.9 (0%)	5.4 (3%)	8.5 (0%)	6.6 (1%)
			Constant	Drift	Mutation	**Epistasis**	Selection	Canalization
T=1	G=10	N=20	0.0 (95%)	3.0 (0%)	6.1 (0%)	5.1 (0%)	6.2 (0%)	5.0 (5%)
		N=100	0.0 (82%)	3.0 (2%)	6.1 (0%)	4.4 (3%)	6.2 (0%)	5.3 (3%)
	G=30	N=20	0.0 (52%)	4.2 (0%)	6.7 (2%)	0.2 (8%)	2.6 (0%)	0.5 (38%)
		N=100	34.2 (0%)	38.4 (0%)	32.5 (8%)	0.0 (35%)	9.0 (6%)	4.9 (51%)
T=2	G=10	N=20	0.0 (97%)	3.6 (0%)	7.4 (0%)	6.0 (3%)	7.6 (0%)	6.4 (0%)
		N=100	0.0 (74%)	3.7 (0%)	7.5 (0%)	2.5 (21%)	7.5 (0%)	6.0 (5%)
	G=30	N=20	103.7 (0%)	108.6 (0%)	113.1 (0%)	0.0 (98%)	14.8 (2%)	107.4 (0%)
		N=100	573.5 (0%)	578.8 (0%)	582.9 (0%)	0.0(100%)	70.8 (0%)	561.7 (0%)

Supplementary Results 4

Impact of genetic drift on parameter estimates. A constant-variance genetic architecture model (equation (6): $\mu_0 = 0.04$, $\sigma^2_A = 0.0001$, $\sigma^2_E = 0.0004$) was implemented in an individual-based simulation. Selection response was simulated for 30 generations in populations which size varied between N=20 and N=1000. Each generation, 0.4N individuals were selected according to their phenotype (the 40% highest phenotypes), as described in the Methods section. 500 simulations were carried on for each population size, and the table below reports the bias of the parameter estimates, as well as the frequency at which the estimated 95% confidence interval contained the true value.

	$(\bar{\hat{\mu}}_0 - \mu_0)/\sigma_P$	$(\bar{\hat{\sigma}}^2_A - \sigma^2_A)/\sigma^2_A$	$(\bar{\hat{\sigma}}^2_E - \sigma^2_E)/\sigma^2_E$
N=20	0.03 (21.4 %)	- 0.22 (6.0 %)	0.10 (27.8 %)
N=50	0.03 (24.0 %)	- 0.10 (9.0 %)	0.05 (61.4 %)
N=100	0.00 (34.8 %)	- 0.05 (22.4 %)	0.02 (87.6 %)
N=200	0.00 (51.0 %)	- 0.03 (27.6 %)	0.01 (99.0 %)
N=500	0.00 (71.4 %)	- 0.01 (47.0 %)	0.01 (100.0 %)
N=1000	- 0.00 (86.2 %)	- 0.00 (64.4 %)	0.01 (100.0 %)

Supplementary Results 5

Comparison between various methods. In order to assess the properties of our framework compared to alternative methods, simulations were run and their output was analyzed with three families of models.

(a) Fixed-effects approach

A constant-variance model was considered as detailed in equation 18.

(b) Random-effect model

The data considered (phenotypic means and variances) do not allow the use of the "animal model" as such, but a random-effect model can still be fit, as described in the methods section (equation 21) and in (Le Rouzic *et al.* 2010).

(c) Least squares regression

The ratio between additive and phenotypic variance (heritability) can be roughly estimated by a least square regression of the realized cumulative selection response on the cumulative selection gradient (forcing the regression to pass through 0). More elaborated methods (e.g. accounting for correlations in the selection responses) exist but were not tested here.

Individual-based simulations were carried on in a purely additive context. During 10 generations, 40% of the $N = 100$ individuals were selected to form the next generation. The initial parameters were $\sigma^2_A = 0.0001$ and $\sigma^2_E = 0.0004$. 1000 simulations were run and the output was analyzed with the three models described above. The bias of the estimate of parameter x is calculated as $(\bar{\hat{x}} - x)/x$, the precision is $\text{sd}(\hat{x})/x$, and the estimated standard error is the average of the standard error calculated from the Fisher information matrix. Only parameters that are directly estimated are presented, with the exception of h^2 for the random-effect model (the software ADMB can estimate the statistical properties of parameter combinations).

All models are somehow biased, but the relative bias remains < 7% in all tested conditions. The simplest regression method severely underestimates h^2, partly because Bulmer effect is not accounted for. Fixed-effect models show satisfactory precision, but underestimate seriously the standard error of the additive variance, which leads to confidence intervals that are too narrow. Random-effect methods seem slightly more biased, but provide very good estimates of the standard error.

σ^2_A	Bias	Precision	Estimated se
Fixed-effects	- 0.04	0.17	0.07
Random-effects	-0.07	0.18	0.20

σ^2_E	Bias	Precision	Estimated se
Fixed-effects	0.01	0.05	0.06
Random-effects	0.01	0.05	0.05

h^2	Bias	Precision	Estimated se
Random-effect	-0.07	0.16	0.16
Regression	- 0.22	0.13	0.03

Supplementary Results 6: Replicated selection responses

It is a common design to repeat selection lines, often twice, and sometimes more. Duplicated selected lines are generated from the same initial population, and are thus expected to share a common initial genetic architecture. Since they are submitted to the same selection pressure, any potential difference between them can be attributed to genetic drift.

1000 simulations were preformed in order to assess whether neglecting drift when analyzing such repeated time series could be problematic. The fixed-effects model was compared with a random-effects setting, which does account explicitly for drift (Supplementary Methods 2). Simulations were individual-based, with $N = 100$ and 40 individuals selected, $\sigma^2_E = 0.0004$, and $\sigma^2_A = 0.0001$ initially. A constant selection gradient ($\beta = 50$) was applied on both lines, and the following table gives the bias, the standard error, and the mean estimated standard error, normalized by the true value of the parameter, for single and duplicated selection experiments, and for both the fixed-effects approach and the random-effect model. The slight bias affects the estimate in the same way for one or two time series. As expected, a random-effect approach provides wider (in this case, more accurate) confidence intervals, but parameter estimates appear to be more biased (especially for duplicated lines) in this implementation. Duplicating time series does not seem to provide more information than e.g. doubling the duration of the series, as already suggested in Le Rouzic *et al.* (2010).

$\hat{\sigma}^2_A$		Bias	Standard error	Estimated standard error
Fixed effects	One ts	- 0.02	0.19	0.07
	Two ts	-0.03	0.12	0.05
Random effects	One ts	- 0.06	0.19	0.20
	Two ts	-0.18	0.11	0.14

Supplementary Results 7

Impact of generation-specific macro-environmental stochasticity on the estimates. A and B: phenotypic times series were obtained from the constant-variance model (equation (6)) and submitted to the 'stochastic simulation' procedure described in the Methods, including macro-environmental effects (independent shifts in the mean phenotype each generation). The average maximum-likelihood estimates were calculated over 500 simulations, for a macro-environmental variance parameter ranging from $\sigma^2_{me} = 0.04 \times 10^{-4} = 1\%\sigma^2_E$ to $\sigma^2_{me} = 8 \times 10^{-4} = 200\%\sigma^2_E$. Parameter values are $\mu_0 = 0.04$, $\sigma^2_A = 0.0001$, $\sigma^2_E = 0.0004$; selection strength is $\beta = 50$. C: the consequences of not estimating the macro-environmental variance were assessed from a more realistic model, the genetic and environmental canalization model (equation (13)), which happened to be the best of the mechanistic models when fit to the fly wing time series. Parameters were: $\sigma^2_{A_1} = 0.000142$, $\sigma^2_{E_1} = 0.000102$, $k_c = 11.9$, $k_g = -3.34$, $\theta = \mu_1 = 0.042$, $N_e = 9.46$. The model was run assuming that the optimum was equal to the initial mean, and the population size was fixed at its true value. Three macro-environmental variances were tested, $\sigma^2_{me} = 0, \sigma^2_{me} = 0.0001, \sigma^2_{me} = 0.0002$, visual inspection of the simulated time series confirms that these figures are higher than what can be usually observed in real time series.

A: No Macro-environmental effects

	$\bar{\hat{\mu}}_0$	$\bar{\hat{\sigma}}^2_A \times 10^{-4}$	$\bar{\hat{\sigma}}^2_E \times 10^{-4}$
$\sigma^2_{me} = 0$	0.0400	0.999	4.054
$\sigma^2_{me} = 0.04 \times 10^{-4}$	0.0399	1.001	4.086
$\sigma^2_{me} = 0.08 \times 10^{-4}$	0.0401	0.992	4.129
$\sigma^2_{me} = 0.2 \times 10^{-4}$	0.0399	0.997	4.228
$\sigma^2_{me} = 0.4 \times 10^{-4}$	0.0401	0.999	4.419
$\sigma^2_{me} = 0.8 \times 10^{-4}$	0.0401	0.992	4.803
$\sigma^2_{me} = 2 \times 10^{-4}$	0.0398	1.001	5.904
$\sigma^2_{me} = 4 \times 10^{-4}$	0.0401	0.996	7.728
$\sigma^2_{me} = 8 \times 10^{-4}$	0.0404	0.995	11.450

B: Macro-environmental effects activated (likelihood equation (25))

	$\bar{\hat{\mu}}_0$	$\bar{\hat{\sigma}}^2_A \times 10^{-4}$	$\bar{\hat{\sigma}}^2_E \times 10^{-4}$	$\bar{\hat{\sigma}}^2_{me} \times 10^{-4}$
$\sigma^2_{me} = 0$	0.0400	1.000	4.037	0.004
$\sigma^2_{me} = 0.04 \times 10^{-4}$	0.0399	1.001	4.045	0.034
$\sigma^2_{me} = 0.08 \times 10^{-4}$	0.0399	1.000	4.042	0.072
$\sigma^2_{me} = 0.2 \times 10^{-4}$	0.0401	0.995	4.046	0.182
$\sigma^2_{me} = 0.4 \times 10^{-4}$	0.0400	1.001	4.050	0.375
$\sigma^2_{me} = 0.8 \times 10^{-4}$	0.0402	0.997	4.056	0.744
$\sigma^2_{me} = 2 \times 10^{-4}$	0.0402	0.997	4.042	1.873
$\sigma^2_{me} = 4 \times 10^{-4}$	0.0399	1.001	4.053	3.760
$\sigma^2_{me} = 8 \times 10^{-4}$	0.0396	1.004	4.042	7.497

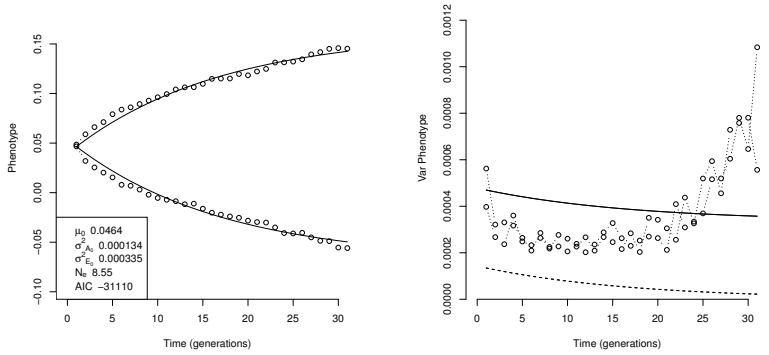
C: No Macro-environmental effects with a canalization model (equation 13))
(mean estimate over 500 simulations $\pm$ standard deviation of the estimate)

	$\bar{\hat{\mu}}_0$	$\bar{\hat{\sigma}}^2_{A_1} \times 10^{-4}$	$\bar{\hat{\sigma}}^2_{E_1} \times 10^{-4}$	$\bar{\hat{k}_g}$	$\bar{\hat{k}_c}$
$\sigma^2_{me} = 0$	0.042	1.423	1.037	-3.394	11.87
	$\pm$ 0.0002	$\pm$ 0.024	$\pm$ 0.083	$\pm$ 0.447	$\pm$ 1.094
$\sigma^2_{me} = 0.0001$	0.042	1.428	1.946	-3.343	7.999
	$\pm$ 0.0012	$\pm$ 0.142	$\pm$ 0.383	$\pm$ 2.565	$\pm$ 2.729
$\sigma^2_{me} = 0.0002$	0.042	1.451	2.867	-3.594	6.369
	$\pm$ 0.0019	$\pm$ 0.228	$\pm$ 0.827	$\pm$ 3.928	$\pm$ 3.993

Supplementary Results 8: macro-environmental effects and 'false' models

Illustration of the potential side effects of macro-environmental variance on the model fitting interpretation. A complex time series was generated, using the parameters estimated from the 11-parameter phenomenological model fit on all experimental *Drosophila* lines (Table 2 and Supplementary Figure 3), and the 'stochastic simulation' procedure described in the 'Methods' section, setting the macro-environmental variance parameter $\sigma^2_{me} = 0$, $N = 100$, and $\beta = \pm 50$. The goal was to obtain a time series in which there is no macro-environmental variance by definition, but too complex to be perfectly fitted by mechanistic models. Drift models (equation (7)) were then fit, without (top panel) and with (bottom panel) macro-environmental variance. Circles represent simulated values, and solid lines illustrate the model predictions. Although the fit is not perfect, the drift model without macro-environmental variance catches the trend of phenotypic means, and provides meaningful parameter estimates. On the contrary, when macro-environmental variance is activated, the best model predicts a high macro-environmental variance, which allows a loose fit. Both models perform as expected when the data is generated from the 'true' model (Supplementary Results 7), and the discrepancy observed here is only due to different ways to handle the misfit. This illustrates a limitation of the macro-environmental effects setting when analyzing real data, due to a situation in which the scientist may prefer a model having a lower likelihood, but providing better predictions for a quantity of interest (the phenotypic mean) at the cost of biased estimates of σ^2_E. Alternatives may include e.g. a penalty for high macro-environmental variances with a prior distribution in a Bayesian setting.

Drift model



Drift + MacroE