Learning from your mistakes: a novel method to predict the response to directional selection

Lisandro Milocco[1,*]

Isaac Salazar-Ciudad[1,2,3,†]

1. Institute of Biotechnology, University of Helsinki, Finland

2. Centre de Recerca Matemàtica, Barcelona, Spain

3. Genomics, Bioinformatics and Evolution. Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, Spain

∗ e-mail: lisandro.milocco@helsinki.fi, ORCID: 0000-0003-3953-0407

† e-mail: isaac.salazar@uab.cat, ORCID: 0000-0003-1607-0962

Manuscript elements: Abstract, Introduction, Methods, Results, Discussion, 5 Main Figures, 2 Appendices, References, 2 Supplementary Figures.

Keywords: quantitative genetics, G-matrix, genotype-phenotype map

Abstract

Predicting how populations respond to selection is a key goal of evolutionary biology. The field of quantitative genetics provides predictions for the response to directional selection through the breeder's equation. However, differences between the observed responses to selection and those predicted by the breeder's equation occur. The sources of these errors include omission of traits under selection, inaccurate estimates of genetic variance, and nonlinearities in the relationship between genetic and phenotypic variation. A key insight from previous research is that the expected value of these prediction errors is often not zero, in which case the predictions are systematically biased. Here, we propose that this prediction bias, rather than being a nuisance, can be used to improve the predictions. We use this to develop a novel method to predict the response to selection, which is built on three key innovations. First, the method predicts change as the breeder's equation plus a bias term. Second, the method combines information from the breeder's equation and from the record of past changes in the mean, to estimate the bias and predict change using a Kalman filter. Third, the parameters of the filter are fitted in each generation using a machine-learning algorithm on the record of past changes. We apply the method to data of an artificial selection experiment of the wing of the fruit fly, as well as to an in silico evolution experiment for teeth. We find that the method outperforms the breeder's equation, and notably provides good predictions even when traits under selection are omitted from the analysis and when additive genetic variance is estimated inaccurately. The proposed method is easy to apply since it only requires recording the mean of the traits over past generations.

Introduction

Evolutionary prediction is an important and active field within evolutionary biology (Lässig et al 2017, Nosil et al. 2018, Shaw 2019, Le Rouzic et al. 2020, Wortel et al. 2021). Aside from its theoretical value, predicting evolution has important applications such as developing strategies for the persistence of populations amidst rapid environmental change (Gomulkiewicz and Shaw 2013, Bonnet et al. 2019), guide the development of vaccines (Hayati et al. 2020) and design interventions to control the spread of a disease (Cobey 2020).

Quantitative genetics is a widely used approach to study and predict short-term evolution of continuous traits (Roff 2007, Walsh and Lynch 2018). The backbone of this theory is the breeder's equation (Lush 1937, Lande 1979, Lande and Arnold 1983). In its multivariate form, it provides predictions of the change in the mean of a set traits, from one generation to the next, in response to directional selection

$$\boldsymbol{\Delta}\bar{\boldsymbol{z}}_i = G_i P_i^{-1} \boldsymbol{s}_i \tag{1}$$

Where $\boldsymbol{\Delta}\bar{\boldsymbol{z}}_i = \bar{\boldsymbol{z}}_{i+1} - \bar{\boldsymbol{z}}_i$ is the vector of change in trait means from generation i to $i+1$, G_i and P_i are additive genetic and phenotypic variance-covariance matrices between traits in generation i, respectively, and $\boldsymbol{s}_i$ is the selection differential in generation i. In this way, the response to selection is predicted as the product of available genetic variation, and a measure selection.

A major appeal of the equation is that its elements can be estimated without detailed knowledge of the genetic architecture and development underlying the focus traits. Indeed, estimates of G_i and P_i can be obtained using only phenotypic data and known genetic relatedness among individuals in a population (Lynch and Walsh 1998, Kruuk 2004), while estimates of $\boldsymbol{s}_i$ need knowledge of individual fitness (Lande and Arnold 1983, Walsh and Lynch 2018). The simplicity of the equation, however, is achieved at the cost of some assumptions.

The breeder's equation assumes an infinitesimal model for genetic effects (i.e. a large number

of loci, each of small effect), or at least a linear parent-offspring regression (and a few additional assumptions, see Rice 2004, Rice 2012, Walsh and Lynch 2018). It further assumes that the population is unselected prior to the application of the equation and that all traits under selection are included in the analysis (Lande and Arnold 1983, Walsh and Lynch 2018, Shaw 2019). Moreover, the equation is local, meaning that the accuracy of the predictions can only be ensured for a single generation (Walsh and Lynch 2018).

When applied to real systems, the assumptions of the breeder's equation are violated to some extent. To start, in practice we only have access to estimates of G_i. This introduces uncertainty and possibly biases to the predictions, particularly when G_i is estimated and used in different environments (Pigliucci 2006) or when relevant effects are not controlled for during the estimation of G_i (e.g. maternal effects, Roff 2007, Pujol et al 2018, Walsh and Lynch 2018). Moreover, the equation is typically used to predict the response for several generations, under the assumption that G_i remains constant over these generations. However, the constancy of the G-matrix is a debated issue (Steppan et al. 2002, Aguirre et al. 2013), and work on nonlinear genotype-phenotype maps (Milocco and Salazar-Ciudad 2021) and gene-environment interactions (Sgrò and Hoffmann 2004, Brodie and Wood 2015) show that the G-matrix can change rapidly even in a few generations. Another common violation is the so-called missing character problem, where the particular traits chosen for study do not account for all selection (Pujol et al. 2018, Shaw 2019).

Indeed, when applied to real systems, violations of the assumptions of the breeder's equation lead to prediction errors (Gimelfarb and Willis 1994, Rice 2004, Roff 2007, Morrissey et al. 2010, Pujol et al 2018, Walsh and Lynch 2018, Shaw 2019, Milocco and Salazar-Ciudad 2020, Pélabon et al. 2021). A notable example is the problem of stasis (Merilä et al. 2001, Shaw 2019) where no response to selection is observed in a population that both has ample additive genetic variance and is under strong directional selection. Prediction errors have also been reported in artificial selection experiments when the parent-offspring regression is nonlinear (Gimelfarb and Willis 1994, Heywood 2005), and when selection is applied in the direction opposite to the sign of the

genetic correlation between two traits under selection (reviewed in Roff 2007).

An important feature of the prediction errors when using the breeder's equation is that their mean over time can be nonzero (Rice 2004, Milocco and Salazar-Ciudad 2020), indicating the presence of a systematic bias. For example, if a trait under selection is missing from the analysis, the prediction using the breeder's equation can be biased because there is an indirect effect of selection that is systematically omitted in the prediction (Merilä et al. 2001). Moreover, if the G-matrix has changed rapidly, for example because the local genotype-phenotype map has a different structure (Milocco and Salazar-Ciudad 2021) or because of enviromental interactions (Wood and Brodie 2015), predictions will also be biased because the G used for predictions is incorrect. The presence of a systematic bias in the predictions means that the error is not purely stochastic, but somewhat structured. In other words, the error at a given generation i is informative of the error at generation $i + 1$. This indicates that there is potential to improve predictions by incorporating this bias, if one could retain the information of past generations as a "memory".

Here, we propose a new method to predict the response to directional selection that yields better predictions when some of the assumptions of the breeder's equation do not hold. The method uses the record of the means of the traits in past generations to improve predictions. There are three key innovations in the method. First, it uses a model for the change in the mean of the traits that is the breeder's equation plus a bias term, which is the term with memory. Second, the method predicts the change in the traits and the bias in each generation using a Kalman filter (Kalman 1960). The filter integrates the information of the breeder's equation and the record of past means of the traits, and it efficiently deals with the noise in the data. Third, the method incorporates a machine-learning scheme to learn the parameters of the filter that provide the best predictions in each generation. Notably, if the assumptions of the breeder's equations are met, the new method reduces to the breeder's equation.

The Kalman filter is a hallmark of control theory (Kalman 1960, Åström and Wittenmark 1997) and has a wide variety of technological applications, from navigation of aircrafts (Grewal and

Andrews 2010) to econometrics (Ghysels and Marcellino 2018). The filter is a general algorithm that allows to estimate the value of a set of variables of interest, using a model of how the variables are expected to change, and a series of measurements observed over time. Here, we adapt it to be used in the prediction of the response to selection.

In the Methods section, we develop the novel prediction method in three parts. Part 1 is the introduction of the extended equation that consists of the breeder's equation plus a bias term. Part 2 is the development of the Kalman filter for this application. Part 3 is the explanation of the machine-learning algorithm to learn the parameters of the filter at each generation. In the Results, the new method is used to predict the response to selection in two artificial selection experiments, and is compared with the predictions using the breeder's equation. We show that the novel method improves the predictions of the response to selection when compared to the breeder's equation, on average. The data sets are used to explore common situations where the assumptions of the breeder's equation are violated to some extent, including when some of the traits that are under selection are omitted from the analysis, when the G-matrix used is outdated, and for varying degrees of precision in the estimation of G_i and P_i. Importantly, these improvements are achieved only by exploiting the registry of past values for the means of the traits. This data is easy to collect, especially when compared to alternatives such as increasing the precision in the estimate of G_i.

Methods

Part 1: The breeder's equation plus a bias term

We want to predict the change in the mean of a set of traits between generations, $\boldsymbol{\Delta \bar{z}}_i$. We propose the following equation consisting of the breeder's equation plus a bias term, $\boldsymbol{b}_i$, a vector of length equal to the number of traits:

$$\boldsymbol{\Delta \bar{z}}_i = G_i P_i^{-1} \boldsymbol{s}_i + \boldsymbol{b}_i \tag{2}$$

The bias term can be understood as the part of the response to selection that is not captured by the breeder's equation, and arises from violation of assumptions, such as presence of nonlinearity in the relationship between genotype and phenotype or missing traits. As such, the systematic bias is structured and we expect the bias of generation i to be similar to the bias at generation $i+1$ (Rice 2004, Milocco and Salazar-Ciudad 2020).

Here we propose to estimate the bias term by using measurements of the system up-to generation i. In principle, one could estimate the $\boldsymbol{b}_{i-1}$ as the difference between the prediction from the breeder's equation, $G_{i-1}P_{i-1}^{-1}\boldsymbol{s}_{i-1}$, and the realized change in the mean, $\boldsymbol{\Delta}\bar{\boldsymbol{z}}_{i-1}$. Assuming that the bias changes slowly, one could further assume that $\boldsymbol{b}_i \approx \boldsymbol{b}_{i-1}$ and obtain an estimation for the bias at generation i. The problem with this approach is that both the breeder's prediction and the change in the mean for the trait are measured with noise, which typically is very large. Then, the estimate of obtained like this would be very inaccurate.

To deal with the problem of noise in the measurements, we propose here to use a Kalman filter to estimate $\boldsymbol{\Delta}\bar{\boldsymbol{z}}_i$ and $\boldsymbol{b}_i$ in each generation. The filter is explained in the next section. To simplify the bookkeeping and notation, we will develop the equations of the filter for each trait separately. We then rewrite equation (2) for each trait as

$$\Delta_i = \Delta_i^B + b_i \tag{3}$$

Where Δ_i is the change in the mean of a given trait in generation i, Δ_i^B is the prediction using the breeder's equation, and b_i is the bias. We want to estimate Δ_i and b_i which we call the state variables.

Part 2: The Kalman filter

The Kalman filter is a general algorithm that integrates two sources of information (Åström and Wittenmark 1997). First, it uses a model of how we expect the state variables to change from one generation to the next. This makes the algorithm recursive, since the estimates of the state

variables at time $i-1$ are used to make estimates of the state variables at i. The information from the estimates at $i-1$ is combined with a second source of information to make estimates of the state variables at time i. This second source of information is a set of measurements from the system, taken at time i and that are related to the state variables. The filter combines these two sources of information by a weighted average. How the average is obtained is the central part of the filter, and it is achieved by calculating a weight matrix that minimizes the error in the estimates (Åström and Wittenmark 1997). Note that both sources of information described above have associated noise, summarized by the covariance matrices R_i and Q_i (explained below). These matrices are the parameters of the filter that have to be provided by the user (see Part 3).

For this particular application of the Kalman filter, the state variables are Δ_i and b_i and they are related to each other by equation (3). Note that with the above definitions, estimating Δ_i gives us a prediction for $\bar{z}_{i+1}$, since $\bar{z}_{i+1} = \bar{z}_i + \Delta_i$. In developing the algorithm below, we use the symbol $\hat{}$ to refer to estimates of the variables (e.g.. $\hat{\Delta}_i$ is the estimate of the state variable Δ_i). We make the usual assumption that the response to directional selection does not show abrupt changes from one generation to the next (Walsh and Lynch 2018). Additionally, in this application we will assume that the bias changes slowly in time. In this way, $\Delta_i = \Delta_{i-1} + \eta_i$ and $b_i = b_{i-1} + \eta_i^b$, where $\boldsymbol{\eta}_i = (\eta_i, \eta_i^b)$ is a vector of small changes that we assume to be normally distributed with mean zero and covariance matrix Q_i.

There are two measurements at time i that we can use to improve our estimates. We use the symbol $\tilde{}$ to indicate that the variable has been measured with noise. The measurements are $\tilde{\Delta}_i^B = \Delta_i^B + v_i^B$ and $\tilde{\Delta}_{i-1} = \Delta_{i-1} + v_i$, where we assume that $\boldsymbol{v}_i = (v_i^B, v_i)$ is a vector of gaussian measurement error with mean zero and covariance matrix R_i.

The Kalman filter combines the estimates of the state variables in $i-1$ (i.e. $\hat{\Delta}_{i-1}$ and $\hat{b}_{i-1}$) and the new measurements (i.e. $\tilde{\Delta}_i^B$ and $\tilde{\Delta}_{i-1}$) to provide the best possible estimates of the state variables in generation i (i.e. $\hat{\Delta}_i$ and $\hat{b}_i$). Given the relationships described above, this is done using the following formula:

$$\begin{pmatrix} \hat{\Delta}_i \\ \hat{b}_i \end{pmatrix} = \begin{pmatrix} \hat{\Delta}_{i-1} \\ \hat{b}_{i-1} \end{pmatrix} + K_i \left(\begin{pmatrix} \tilde{\Delta}_i^B \\ \tilde{\Delta}_{i-1} \end{pmatrix} - \begin{pmatrix} \hat{\Delta}_{i-1} - \hat{b}_{i-1} \\ \hat{\Delta}_{i-1} \end{pmatrix} \right) \tag{4}$$

The first term of the right-hand side is the state vector estimates in step $i-1$. The second term is the correction, which is the product of the matrix K_i and the error. The error is formed by the difference between the measurements $\tilde{\Delta}_i^B$ and $\tilde{\Delta}_{i-1}$, and their expected values using the estimates at step $i-1$.

K_i is a 2×2 matrix called the Kalman gain, which assigns weights to the correction. The calculation of K_i is the key of the filter, and it is done for each i. K_i is a trade off between the confidence we have on the estimate of the states at $i-1$ and the confidence we have on our measurements at generation i, and is calculated to minimize the error covariance of the estimates (Kalman 1960, Åström and Wittenmark 1997). If the measurements are to be trusted, then the gain will give more weight to the second term of equation (4). If the estimates at $i-1$ are to be trusted, then the gain will assign more weight to the first term of the equation. The "trust" is quantified by the associated error covariance matrices. This, together with the calculation of the gain K_i is explained in the Appendix A.

As mentioned above, the algorithm is recursive: the estimates obtained in generation $i-1$ using equation (4) is the starting point for the prediction in generation i. We then require initial estimates at time $i = 0$ to begin the recursion. For our state variables, $b_0 = 0$ and $\Delta\bar{z}_0$ is the prediction using the breeder's equation.

Part 3: Learning the parameters of the Kalman filter

The matrices Q_i and R_i have to be provided by the user to implement the filter explained in Part 2 of the method. Q_i is the covariance of the vector $\boldsymbol{\eta}_i$, and R_i is the covariance of vector $\boldsymbol{v}_i$, which describes measurement noise. These matrices are hard to calculate analytically. For example, the variance in the measurement noise for $\tilde{\Delta}_{i-1}$ is affected by drift, selection, measurement and

sampling (Walsh and Lynch 2018). An added difficulty is that R_i and Q_i can change in time.

Instead of calculating the matrices R_i and Q_i analytically, we learn them using the time series of the trait means. That is, at generation i we use a window of the last L recorded changes in the mean $\{\tilde{\Delta}_{i-L}, ..., \tilde{\Delta}_{i-1}\}$ to learn the values of R_i and Q_i. This is done by running the filter inside the window with several combinations of R_i and Q_i . We then calculate the prediction error of the method in the window for each combination of R_i and Q_i , and keep the combination that results in smallest prediction error. We then use this combination of R_i and Q_i to make the actual prediction of interest at time i. Note that this process is done in every generation i for each time series separately. In this way, the method learns the best R_i and Q_i possible for the specific system at time i.

To learn the matrices, we assume that R_i and Q_i are roughly constant inside the window. This sets a limit to how large the window can be, since if the window is too large then the matrices may change substantially inside the window. Then, the size of the window should be kept relatively small, making it hard to learn all the elements of the 2×2 matrices R_i and Q_i (i.e. more elements to learn require a larger dataset). To reduce the number of elements to learn, we make the additional simplification that R_i and Q_i are diagonal, and that each has the same elements in the diagonal. This simplification is reasonable because both state variables have similar magnitudes, as they are both related to change of the same trait. Details of how this is done are given in Appendix B. For the analyses in this paper, we use a window of size $L = 5$ for $i > 5$, and $L = i$ for $i \leqslant 5$ (i.e. we use the available generations in the record).

Apart from using the window to learn R_i and Q_i, we also use it to approximate the uncertainty in the predictions using the new method. To do this, we calculate the standard deviation of the residuals of the predictions against the observed change inside the window. We use this as the uncertainty for the predictions using our method.

The artificial selection experiments

We use data from two selection experiments to test the new method and the breeder's equation in their ability to predict evolutionary change. Experiment 1 is simulated data for teeth evolution, and Experiment 2 is data from an experiment in the wing of the fruit fly.

Experiment 1: teeth

We use data of an in-silico artificial selection experiments on teeth. Details of the simulations are given in previous work, and the data is publicly available (Milocco and Salazar-Ciudad 2020, Milocco and Salazar-Ciudad 2021). Briefly, each evolutionary simulation has a population of genotypes. Each genotype is mapped to a tooth morphology through a deterministic model of tooth development (Salazar-Ciudad and Jernvall 2010, Harjunmaa et al. 2014). The tooth model recapitulates the process of development for a tooth, starting from a flat epithelium to a complex 3D morphology. The dynamics of development, and the resulting phenotype, are determined by the value of a set of parameters which are determined by the genotype. This means that variation in genotypes results in variation in the phenotypes, but the mapping between these types of variation is complex and ultimately determined by the tooth development model itself. Traits were measured on each teeth. These were the x- and y-coordinates of 3 landmarks located in the 3 tallest cusps of the tooth (see Figure 1A). In each generation, once the genotypes of all individuals had been mapped to their corresponding phenotypes using the tooth development model, selection was applied by choosing 50% of the individuals with morphology closest to the optimum. Each simulation had an optimum shape, defined at the beginning, which determined the direction of selection (see Figure 1C). Selected parents were paired randomly, and produced the next generation of genotypes. Each couple produced 4 offspring, resulting in a constant population size. Recombination and mutation were included in each generation, and the process was iterated to simulate evolution. There is a total of 32 simulations, each with a different selection optimum. Each simulation was run for 30 generations using a population of 300 males

and 300 females.

Estimation of variance components and observed change: In each generation, the elements of the breeder's equation were estimated (i.e. G_i, P_i. and $\boldsymbol{s}_i$). Variance components were estimated from a half-sibling breeding design using individuals at generation i as base population (details in Milocco and Salazar-Ciudad 2020). The animal model used was the simplest possible (i.e. with only additive genetic merit fitted to each individual). Restricted maximum likelihood (REML) estimates of G_1 and P_1 were obtained using the software WOMBAT (Meyer 2007). Sampling variation in the estimation of G_1 was accounted for using the REML-MVN method (Houle and Meyer 2015). The method approximates the uncertainty in evolutionary parameters estimated using animal models by resampling G-matrices from the distribution of its maximum-likelihood estimate. For each generation, we resampled 100 G- and P-matrices from this distribution and used them to calculate 100 predicted changes using the selection differential and the breeder's equation. We plot the mean and 1 SD of these predictions. Note that the tooth development model is deterministic and there is no measurement error. Moreover, we have a large sample size. This allows for very precise estimates of G_i and P_i.

Due to the fact that there is little measurement noise for the population mean in the simulations, the *observed change* was obtained directly as $\Delta_i = \bar{z}_{i+1} - \bar{z}_i$. This is the amount that we look to predict at generation i (see Figure 1E, G).

Experiment 2: fruit fly wing

We performed artificial selection experiments on the wing of the fruit fly *Drosophila melanogaster*. The starting population was founded from 250 isofemale lines derived from flies captured during the Summer of 2017 in Groningen, The Netherlands by the Billeter's lab. From each line, 25 females and males were collected and merged to make a large, outbred population that was maintained in laboratory conditions. For the initial generation of the experiments, 100 virgin males and 100 virgin females from the large population were randomly assigned to one of four lines. Three of these lines were subjected to selection (R1, R2 and R3), with the remaining being

a control without selection (C1). Lines were kept at 25°C with alternating 12-h light and dark cycles during the experiment.

In each generation, 100 males and 100 females were collected as virgins. The left wing of each collected, anesthetized fly was taken by the automatic system known as the WingMachine (Houle et al. 2003, Mezey and Houle 2005). The x- and y-coordinates of 5 landmarks shown in Figure 1B were obtained using a semiautomatic landmarking software (see Houle et al. 2003 for details, note that we use a subset of 5 landmarks from the total 12 landmarks provided by the pipeline). In the Control line, 50 males and 50 females were chosen randomly as parents for the next generation. In the selected lines, the 50 males and 50 females with wings with shortest distance to the optimum morphology were selected as parents. The distance of each individual to the optimum was calculated as the euclidean distance between the values of the traits in the individuals and the optimal values of the traits. The optimum morphology is shown in Figure 1D, and is the same for the three lines with selection. The process of image processing and selection was repeated in each generation. Sibling mating was avoided to reduce inbreeding. The process was repeated for a total of 20 generations, equivalent to 4000 flies per line (16 000 in total). If some of the formed couples did not produce offspring for the next generation, either because one of the parents died or due to infertility, we measured more offspring from other couples to complete the 200 individuals per generation. We also formed 3 extra couples in each generation, to provide extra individuals in case some of the original 50 couples failed to produce offspring.

As mentioned above, we measured the x- and y-coordinates of 5 landmarks, resulting in 10 traits. The data was aligned by generalized Procrustes least squares superimposition. Four degrees of freedom are lost in this process, one to estimate wing size and three to standardize the orientation of wing shapes. Therefore, there are only 6 independent traits in the data. For these traits to be comparable between lines and through the generations, we use the 6 first components of a PCA of generation 1 of the Control as a reference and project the all data to that space. The resulting 6 phenotypic traits are a linear combination of the original 10 traits that conserves all

relevant variation in all lines. In this paper, we refer to these 6 traits as the phenotypic traits. The means of these traits agains generations, for all 4 experimental lines, are shown in Supplementary Figure 1.

Estimation of variance components and observed change: All lines start from the same founding population. We estimate G_1 and P_1 for this founding population by pooling the first n generations of the Control. For these n generations we have the pedigree and phenotypic data. We call n the depth of the pedigree. Here we explore values of n from 2 to 15. Larger n will result in more accurate estimates of G_1 and P_1, but requires more measurements. We also compared the predictions of the breeder's equation and the new method in the case that $G_1 = P_1$. REML estimates of G_1 and P_1 were obtained using the software WOMBAT (Meyer 2007) and sampling variation was estimated using the REML-MVN method (Houle and Meyer 2015). The animal model used included sex, generation and ID of the person measuring as fixed effects.

The estimation of the means in each generation inevitably has noise. Noise arises from the imaging and landmarking process, finite sampling of the population and drift. Because we focus on directional selection, this noise has to be removed. We perform a quadratic regression to the 20-generation time series of the means, which is a common regression for long-term artificial selection data (Eisen 1972, Rutledge et al. 1973, Grassini et al. 2013, Walsh and Lynch 2018). The fitted values are used as Δ_i, which we call the *observed change*. This is compared to the change predicted by the new method and the breeder's equation (see Figure 1F, H).

Results

We compare the performance of the new method introduced here and the breeder's equation in predicting the response to directional selection, using two artificial selection experiments summarized in Figure 1. The performance of the prediction methods is assessed by calculating the prediction error, obtained as the relative root mean squared error (RMSE) between the multivariate series of predictions and the multivariate series of observed changes. This is, for a given times

series, all the predicted changes for all generations are stored in a matrix of predictions, and all the observed changes are stored in a matrix of observations. The RMSE is calculated between those matrices as the square root of the sum of the squared differences, divided by the square root of the sum of squared elements of the matrix of observations. This is a general measurement of the goodness of prediction for the whole time series.

Predicting the response to selection in teeth simulations

The teeth artificial selection experiments are in silico simulation of evolution in a population. A key feature of these simulations is that the mapping between genetic and phenotypic variation is done using a model of development that produces realistic morphological variation (Salazar-Ciudad and Jernvall 2010). Importantly, the genotype-phenotype map of this model is known to be complex and lead to biases in the estimation of the response to selection (Milocco and Salazar-Ciudad 2020). There are a total of 32 simulations with different optima, each of 30 generations (Figure 1C). Figure 1 shows the tooth morphology and the 3 landmarks used. The x- and y-coordinates of these landmarks are the 5 measured traits. Figure 1 also shows the response to selection for one trait in an example simulation. Because the data is simulated, all conditions are controlled. This allows to isolate specific sources of prediction error, and test how well the new method is able to perform. Specifically, we test the situation where the G-matrix is outdated by a varying number of generations, and when traits that are under selection are omitted from the prediction.

First, we study the scenario where the P- and G-matrices are not known in all generations. This is the most common scenario, since obtaining estimates for each generation, or for blocks of generations, is very expensive. We update the estimates of P_i and G_i every a given number of generations that we call the update time. For example, for an update time of 10, the matrices are calculated at generation 1, 11 and 21 of each experiment. We use update times of 8, 10, 15 and 30 generations, which correspond to 4, 3, 2 and 1 samples of G_i throughout the experiment, respectively.

Figure 2F summarizes the prediction error for all simulations, using the breeder's equation and the new method for different update times. Each point in the scatter plot is the error measured as RMSE between the multivariate series of predictions and the multivariate series of observed changes, for the new method and the breeder's equation. The error is significantly smaller using the new method, and the improvement is more clear for larger update times (i.e. when G_i is updated with less frequency).

Figure 2A-E show the time series of observed and predicted changes for example simulation 18 using a time update of 30 generations. In this simulations, the population crosses a region of the genotype-phenotype map that is nonlinear (Milocco and Salazar-Ciudad 2020). This results in relatively fast changes in the observed response to selection (see for example panel B from generations 5 to 15). The new method provides much better predictions than the breeder's equation in those regions.

We use the teeth artificial selection experiments to study the situation where traits that are under selection are omitted from the prediction, what is known as the missing character problem. To isolate this error and avoid confounding it with the error arising from using old estimates of G_i and P_i, we use estimates of variance components at each generation (i.e. update time of 1). Note that the error would increase if we used an update time larger than 1. We try to predict the change in traits 2 and 3 (i.e. the x- and y- position of the landmark located in the posterior cusp, see Figure 1) without data from traits 1, 4 and 5. Figure 3A-B shows the predicted and observed changes for traits 2 and 3 in an example simulation. We find that this omission can lead to biases, and that the new method is able to correct the errors to a large extent. A summary for all simulations is given in Figure 3C.

Predicting the response to selection in the wing

The artificial selection experiment in the wing shows the full complexity of the problem of predicting the response to selection in a real population. This is the most common scenario in which the new method can be applied. There are three replicates with selection and one control line,

all coming from the same base population and running for 20 generations (see Methods and Supplementary Figure 1). In each generation, 100 males and 100 females are measured. Selection is applied on 5 landmarks of the wing as shown in Figure 1B, by selecting the 50% of measured individuals in the direction shown in Figure 1D.

For this experiment, we only calculate the G-matrix at the beginning (i.e. G_1). Since the control line and the selection lines all start from the same base population, we use the pedigree and phenotypic data of the initial generations of the control to estimate G_1. We call the pedigree depth the number of generations of the control line used to estimate G_1. The larger the pedigree depth, the more precise the estimate of G_1. We test the prediction ability of the breeder's equation and the new method using estimates of G_1 for different pedigree depths, ranging from 2 to 15 which correspond to 400 to 3000 individuals from the control.

Figure 4 shows the predictions for the change in the traits using the new method and the breeder's equation, against the observed change, for selection line R2 (other lines shown in Supplementary Figures 2). A pedigree depth of 2 was used here. It can be seen that the new method yields predictions that are closer to the observed change, particularly for traits 1, 3 and 6. Also note that the change for trait 4 is accurately predicted by the breeder's equation. In this case, the new method performs as well as the breeder's equation.

Figure 5A shows the prediction error for the new method and the breeder's equation for different pedigree depths. The figure shows that the new method outperforms the breeder's equation for all pedigree depths. The plot includes a pedigree depth of 1, which means assuming that $G_1 = P_1$, i.e. that all phenotypic variation is additive genetic. Notably, the new method using a G-matrix with small pedigree depth outperforms the breeder's equation using a G-matrix with a large pedigree depth. The most extreme case is in R3, shown with triangles, where the predictions using $G_1 = P_1$ and the new method are better than the predictions using the breeder's equation and a very precise estimate of G_1. This is important because, experimentally, it is much more expensive to increase the accuracy of the estimate of G_1 that to apply the new method. The latter only requires recording the trait means in past generations, while the former

requires phenotypic and relatedness data in particular breeding designs.

We found that there are two sources of prediction error for the breeder's equation in these experiments, and that the new method provides significantly better predictions in both cases. First, there is error associated to using wrong estimates of G_1 and P_1. This prediction error is most evident when using estimates of G_1 with low pedigree depth, and even more when assuming $G_1 = P_1$. Increasing the pedigree depth of the estimates can correct much of these errors. The second source of error is the fact that G_i changes during the experiment. This leads to possible errors at later stages of the experiment if the G-matrix estimated for the base population is used, even if the estimate of G_1 is obtained with high accuracy (i.e. a deep pedigree in this case). Both of these error are shown in Figure 5B for trait 6 of line R1, using different pedigree depths, namely 1 ($G = P$), 2 and 10. A big part of the error is reduced when increasing the pedigree depth. However, even when using a precise estimate of G_1 (pedigree depth of 10), the breeder's predictions remains significantly biased towards the end of the experiment. Regardless of the source of the error, the novel method outperforms the breeder's equation as shown in Fig. 5B.

Discussion

We developed a novel method to predict the response to directional selection by combining the breeder's equation with data from the time series. We tested the new method with two artificial selection experiments, and show that it outperforms the breeder's equation. The method is general, and can be applied to virtually any evolving system under directional selection. Most importantly, the new method only requires the record of means of the trait for past generations, which is relatively easy to collect, at least compared to alternatives like obtaining better estimates of G_i. An important feature of the new method is that it reduces to the breeder's equation when the assumptions of the latter are met (that is, when b_i=0 and R_i=0). In this way, the method can be applied to a wide variety of scenarios, specially where the assumptions of the breeder's

equation are not met, like in the later stages of long-term selection studies, when the full set of traits under selection is not known and when G_i cannot be accurately estimated. The more the assumptions are violated, the more the new method will outperform the breeder's equation, as shown in Figure 2F and 5A.

We highlight three key aspects of the novel method. The first aspect is the introduction of the bias term in equation (2). This is proposed on the grounds of previous theoretical and empirical work that shows that the expected value of the prediction error using the breeder's equation may not be zero (Rice 2004, Pujol et al 2018, Walsh and Lynch 2018, Milocco and Salazar-Ciudad 2020). This indicates that the error at generation $i-1$ is correlated with the error at generation i. The second key aspect of the method is the use of a Kalman filter. This was necessary to deal with the noise associated with the measurements of the time series and the breeder's equation.

The final key element of the method is that we use a window of data to learn the parameters of the filter for each generation using a machine-learning algorithm. This is a significant conceptual shift from classical quantitative genetics approaches which are "offline", meaning that parameter estimation is done after all the data has been collected. For example, the G- and P-matrices are obtained offline with data from a given a population, and later used to make predictions on the same or a different population. Similarly, realized heritabilities are calculated offline from regression of the response after the selection experiment is completed, commonly pooling information from different replicate lines (Walsh and Lynch 2018). The offline approach underexploits the dynamical properties of the time series (Le Rouzic et al 2011), as it does not capture possible temporal changes of the parameters, and other singularities of the time series. The method we propose here, on the contrary, works "online" by calculating the parameters dynamically at each time point and exploiting the information in the time series data. Apart from enabling the method to be used in real time (i.e. during the experiment), it has the important quality that it allows the parameters to change in time. Moreover, it uses information that is specific to the population of interest and its singularities. This avoids extrapolating information taken in different conditions, which is a known problem in quantitative genetics (Pujol et al 2018).

The method introduced here can be classified as recursive, because it forecasts the variables of interest using past estimates of the variables. There has been recent interest in recursive models to make predictions of future evolution (Le Rouzic et al 2011, Nosil et al. 2018, Rescan et al. 2020, Nosil et al. 2020, Rescan et al. 2021). For example, Nosil et al. 2018 fitted an autoregressive model using several years of data of frequency changes of a morphs and patterns in a population of stick insects. They examined whether data from early time points in the series could predict data in later time points of the series (similar to what we do using the window of past generations). They were able to successfully predict changes in frequency for a trait under clear frequency-dependent selection, but failed to predict change for a trait under a more complex, unknown form of selection. They conclude that predictability was limited by understanding of selection. The authors suggest that knowledge of selection could be determinant in improving predictions when using recursive models. The method we propose in this paper does exactly this: it combines a recursive model, given by the window of past generations, with knowledge of selection, given by the breeder's equation. Used like this, the breeder's prediction contributes the type of information that purely recursive models are lacking. At the same time, purely theoretical models like the breeder's equation are based on simplifying assumptions that may miss some of the complexity of the system, and work offline. The efficient combination of the recursive model, which is data-driven, and the breeder's equation, which is theoretical, is what results in the method proposed here to outperform each approach when used separately.

The power of the method is best shown in Figure 5. The new method outperforms the breeder's equation regardless of the accuracy in the estimation of the G-matrix, which is the limiting step in applying the breeder's equation. Moreover, the figure shows that the new method using an inaccurate G-matrix is better than the breeder's equation using a very accurate, and expensive to estimate, G-matrix. Even more, the method is able to make big corrections and overall provide very good estimates even when G is not estimated at all. That is, assuming that $G_i = P_i$, which corresponds to precision 1 in the x-axis of Figure 5A. This is an important result because the P-matrix has been used as a proxy of the G-matrix for morphological traits (i.e.

Cheverud's conjecture, Cheverud 1988, Assis et al. 2016, Sodini et al. 2018, Love et al. 2021), a simplification suggested due to the difficulty in estimating the latter. When used in this method, this approximation works because the resulting deviations are corrected by the bias term. Note that even in this case, the information of selection is still exploited, as it enters the predictions through the selection differential, s_i. An important note is that both the breeder's equation and the new method perform significantly better when G_i is estimated than when it is assumed that $G_i = P_i$, (compare precision 1 and 2 in Figure 5A). This means that G_i contains useful information, even when estimated with relatively low precision.

The method proposed here is specific to continuous directional selection, sustained for several generations. This allows to develop the specific equations explained in Part 1 of the Methods. In principle, a similar framework combining multiple sources of information could be developed for other types of selection, such as fluctuating. The difficulty here may be in obtaining information of how selection is acting in each generation. Recent efforts (Rescan et al 2021) have tried to map environmental fluctuations to fluctuations in selection, since certain environmental queues such as temperature are much easier to measure than selection itself. Developing such an approach could allow to improve predictions by measuring environmental queues, and feeding the measurements into the predictive model of a similar form to the one described here.

Data-driven methods are only becoming more popular in the future. This change from more classical, theoretical methods is fueled by the rapid accumulation of data. The method we propose here is line with this change, by combining theory and data. As suggested by other authors (Nosil et al. 2020) this is a promising future for developing better predictions in evolutionary biology. We hope that the method proposed here will be widely applied since it provides better predictions with very few additional costs.

Figures

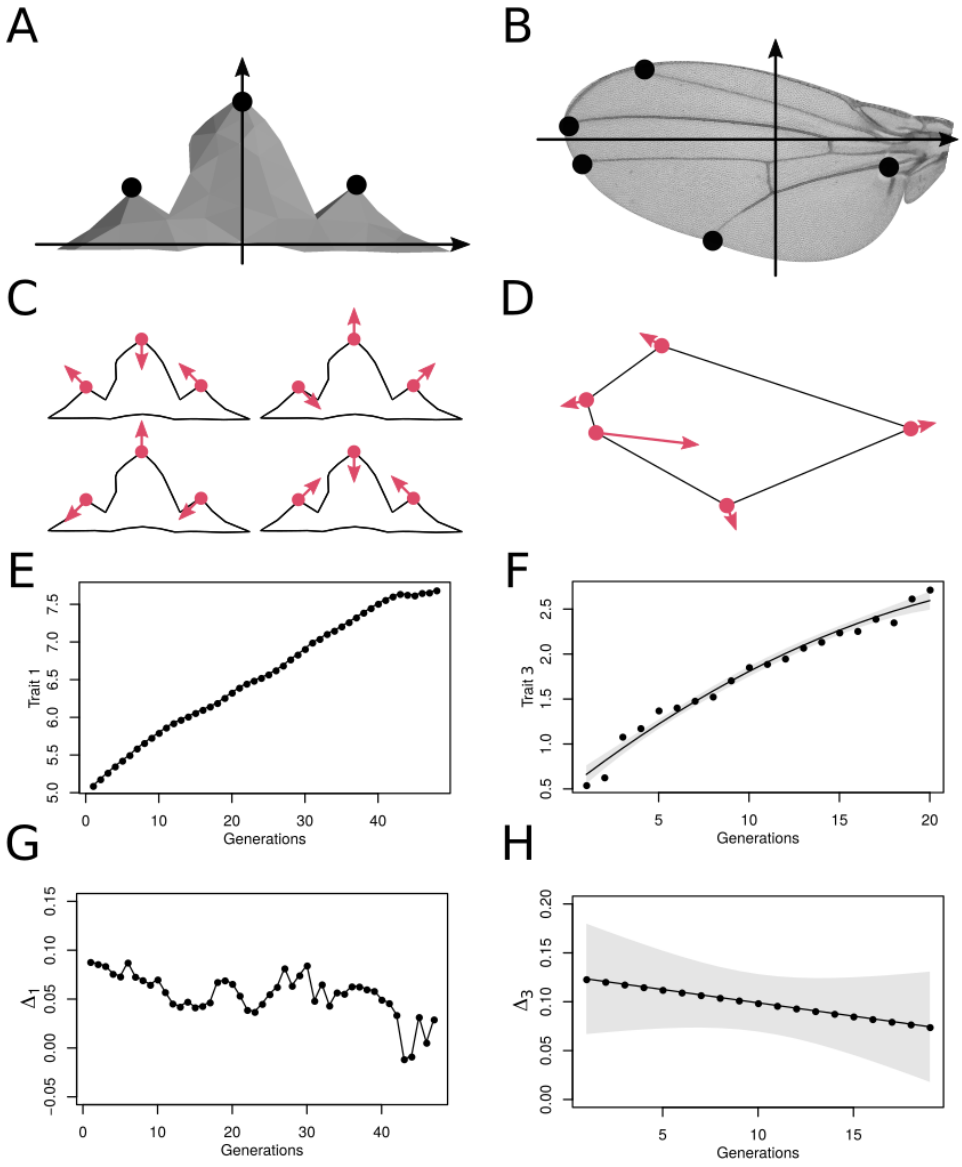


Figure 1: Summary of the artificial selection experiments. A, C, E, G correspond to the teeth experiments and B, D, F, H to the fly wing experiments. A shows the tooth morphology and the 3 landmarks used for the experiments. The coordinates of the landmarks are the phenotypic traits. Note that the y-axis passes through the central landmark, resulting in 5 traits (i.e. x- and

511 y- coordinates for anterior and posterior landmark, and only y-coordinate for central landmark).
512 C shows the directions of selection for 4 of the 32 evolutionary simulations as examples. All
513 combinations of up and down selection for the traits are used. E is the mean of trait 1 in time
514 for simulation 1, and G shows the change in the trait mean. We do not make a regression in E
515 because there is little measurement noise. B shows the morphology and the five landmarks on
516 the wing. There are 6 phenotypic traits that are obtained after aligning the 10 coordinates of these
517 landmarks using Procrustes superimposition. D shows the direction of selection. F shows the
518 mean of trait 3 for line R2, together with a quadratic fit to the data and its 95% CI. The remaining
519 traits and replicate lines are shown in Supplementary Figures 1 and 2. H shows the change in
520 the mean of trait 1 (i.e. what we aim to predict) and its 95% CI.

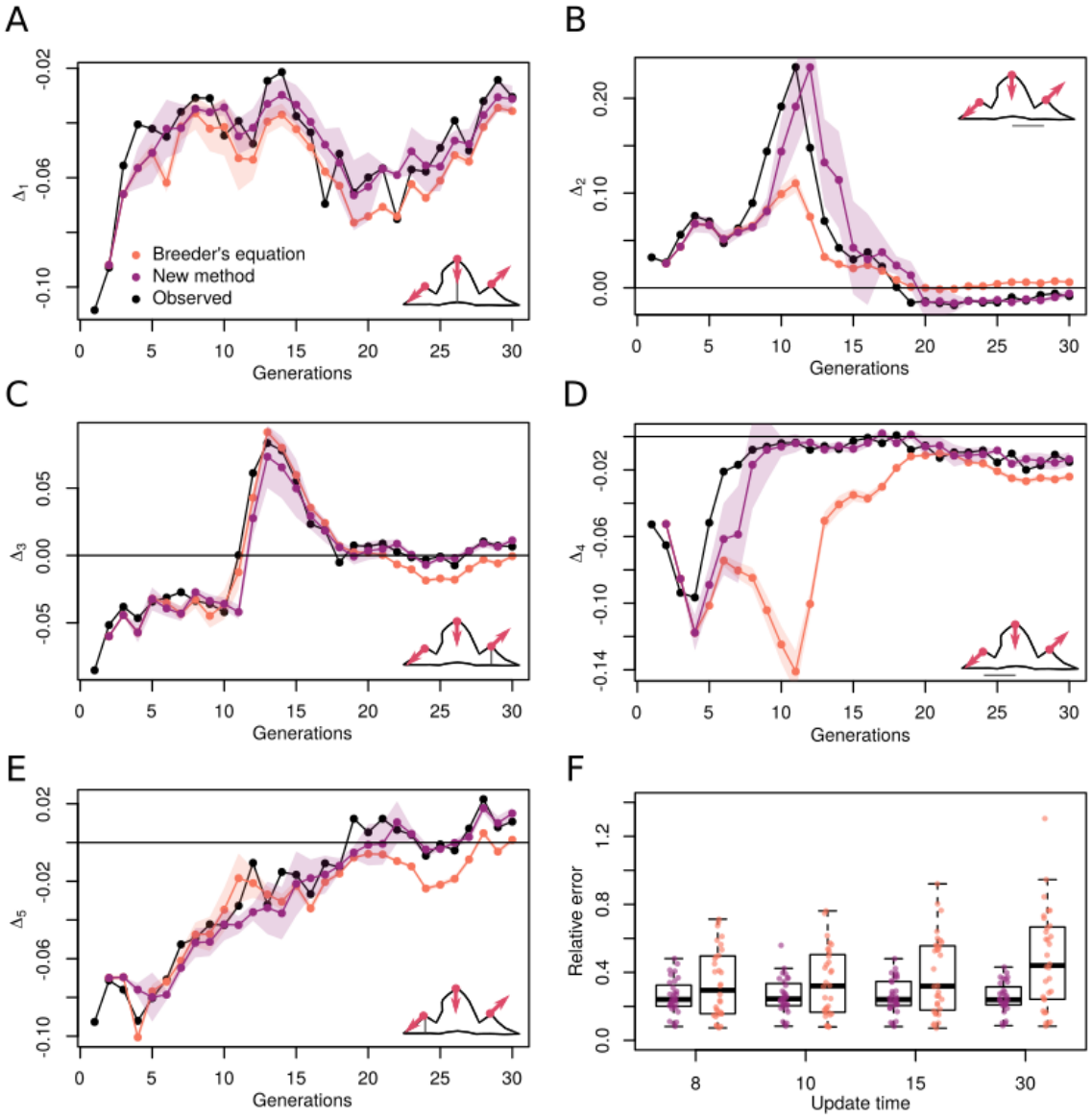


Figure 2: Breeder's equation and the new method applied to predict the response to selection in the teeth experiments. *A–E* are the predictions and obersved changes for all 5 traits of example simulation 18, using the G_1 and P_1 matrices (i.e. estimated at generation 1). Each plot shows a diagram of the tooth, the direction of selection for this particular simulation and a a gray bar indicating the trait being plotted. *F* shows a summary of the prediction error for all simulations, when updating the estimates of G_i and P_i every 8, 10, 15 and 30 generations. For

these experiments of 30 generations, this means 4, 3, 2 and 1 update(s) of the G-matrix throughout the experiment. Each point in the boxplot is the total error for one of the 32 simulations. The new method always has smaller error that the breeder's equation, but the improvement is more clear when G is updated with less frequency.

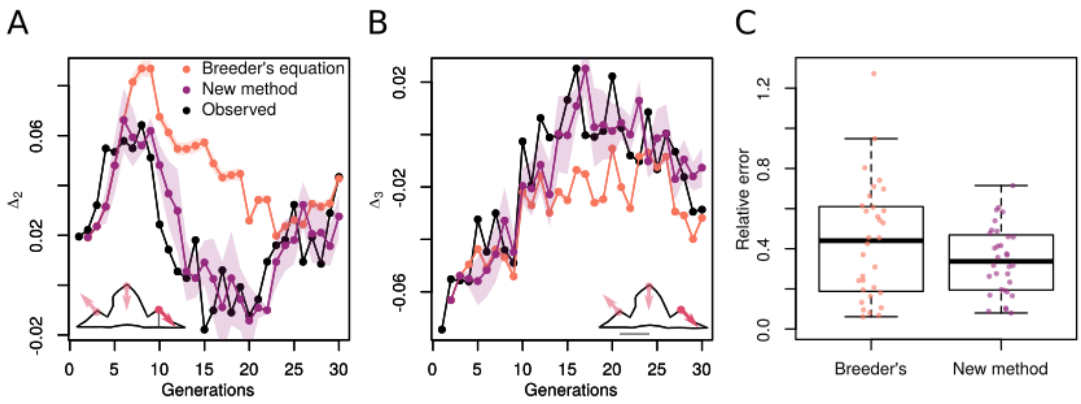


Figure 3: Observed and predicted change in the teeth experiments when omitting traits under selection in the prediction. Only traits 2 and 3 (i.e. the *x*- and *y*- coordinated of the posterior cusp of the tooth) are used for predictions. *A* and *B* show the 2 traits for example simulation 32. The diagram of the tooth inside the panel shows that only the posterior cusp is considered for predictions. *C* shows a summary for all simulations when making predictions using only traits 2 and 3. Each point in the boxplot represents one of the 32 simulations.

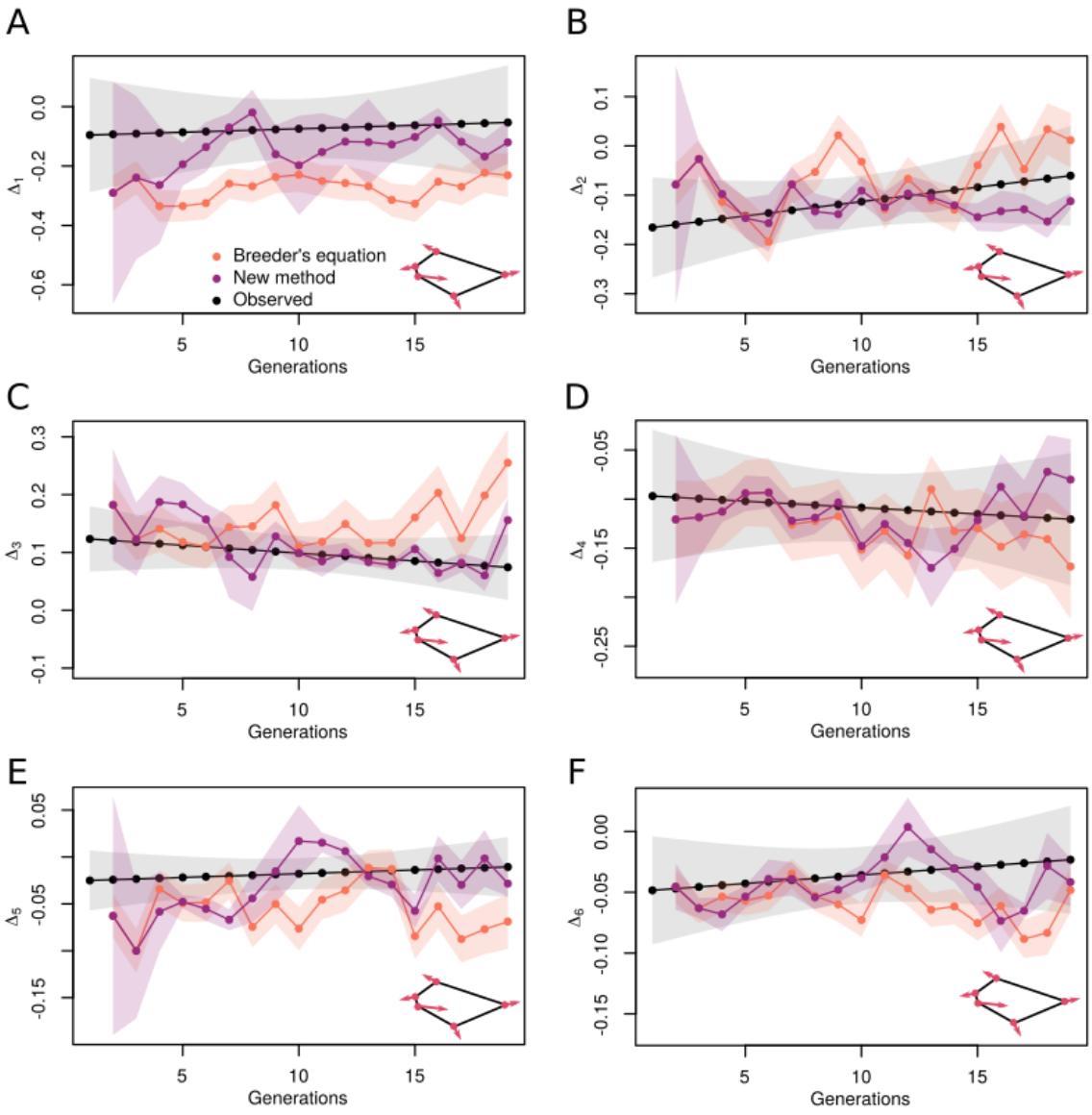


Figure 4: Response to selection for the 6 phenotypic traits in line R2 of artificial selection experiments with the wing. The predicted and observed changes are shown with their approximated uncertainties (95% CI for the observed change, 1 SD for the predictions). The observed change is obtained from a regression of the means (see Fig. 1). A pedigree depth of 2 was used to estimate the variance components.

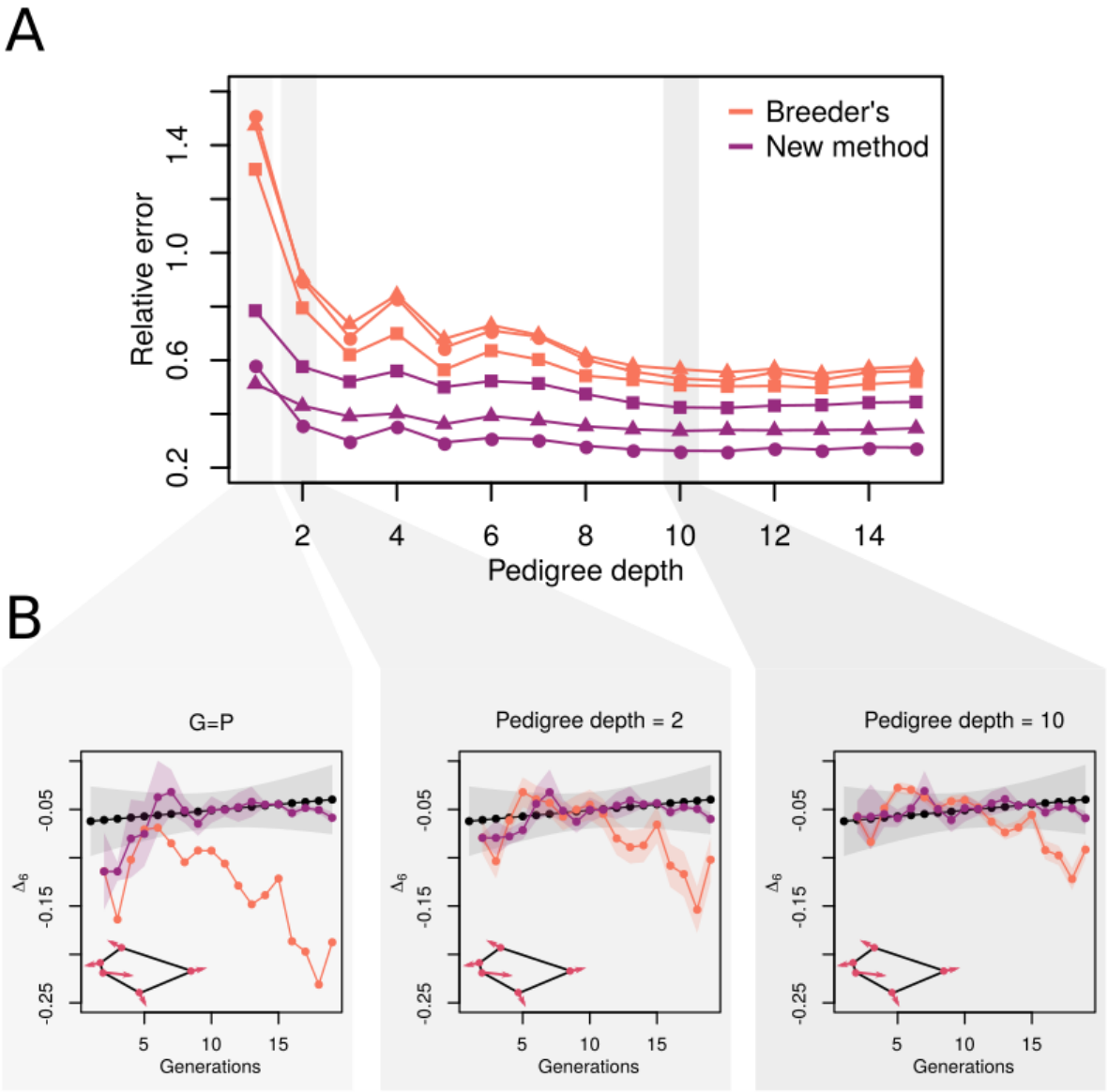


Figure 5: Predicted and observed response to selection for different pedigree depths in the artificial selection experiments of the wing. *A* shows the total error of the prediction using the breeder's equation (orange) and the new method (purple) for the 3 replicated selection lines (squares correspond to replicate R1, circles to R2 and triangles to R3). We performed predictions using the two methods and G_1 and P_1 matrices estimated with varying pedigree depths. We also include the case where we assume $G_1 = P_1$, and this is plotted as pedigree depth of 1. The new

548 method outperforms the breeder's equation for all pedigree depths. B shows the predicted and
549 observed time series for trait 6 of R1, for different pedigree depths. Color code is the same as
550 Figure 2, with the observed change in black. Increasing the pedigree depth of the estimate of G_1
551 reduces the prediction error, but the predictions of the breeder's equation remain biased towards
552 the end of the experiment even with high pedigree depth. The new method is able to correct
553 this. Approximate uncertainties are included as shaded areas. Note that for $G_1 = P_1$ there is no
554 uncertainty for the breeder's equation since variance components are not estimated.

Appendixes

Appendix A

Here we will derive the equations of the Kalman filter, importantly how the gain matrix K_i is calculated in each generation. For this, we first need to obtain an expression for the error which is what we want K_i to minimize. We will use a matrix representation of equation (4),

$$\hat{\boldsymbol{x}}_i = \hat{\boldsymbol{x}}_{i-1} + K_i(\tilde{\boldsymbol{y}}_i - C\hat{\boldsymbol{x}}_{i-1}), \tag{5}$$

where

$$\hat{\boldsymbol{x}}_i = \begin{pmatrix} \hat{\Delta}_i \\ \hat{b}_i \end{pmatrix}, \ \tilde{\boldsymbol{y}}_i = \begin{pmatrix} \tilde{\Delta}_i^B \\ \tilde{\Delta}_{i-1} \end{pmatrix}, \ C = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}. \tag{6}$$

Considering that

$$\boldsymbol{x}_i = \boldsymbol{x}_{i-1} + \eta_i, \text{ where } \boldsymbol{x}_i = \begin{pmatrix} \Delta_i \\ b_i \end{pmatrix}, \tag{7}$$

the measured vector can be rewritten in terms of past state variables as follows:

$$\tilde{\boldsymbol{y}}_i = C\boldsymbol{x}_i + \boldsymbol{v}_i = C(\boldsymbol{x}_{i-1} + \boldsymbol{\eta}_i) + \boldsymbol{v}_i \tag{8}$$

where

$$\boldsymbol{v}_i = \begin{pmatrix} v_i^B \\ v_i \end{pmatrix}, \ \boldsymbol{\eta}_i = \begin{pmatrix} \eta_i \\ \eta_i^b \end{pmatrix} \tag{9}$$

Then, using (7) and replacing (8) in (5), the matrix representation for the prediction error, $\boldsymbol{e}_i$, is

$$\begin{aligned} \boldsymbol{e}_i &= \boldsymbol{x}_i - \hat{\boldsymbol{x}}_i \\ &= \begin{pmatrix} I & -K_i \end{pmatrix} \left[\begin{pmatrix} I \\ C \end{pmatrix} \boldsymbol{e}_{i-1} + \begin{pmatrix} \boldsymbol{\eta}_i \\ C\boldsymbol{\eta}_i + \boldsymbol{v}_{i,} \end{pmatrix} \right] \end{aligned} \tag{10}$$

where I is the identity matrix. The covariance of the error (Φ_i) is the expected value of the product of the error by it transpose, $\Phi_i = \mathcal{E}[\boldsymbol{e}_i \boldsymbol{e}_i^T]$. Taking into account that $\boldsymbol{\eta}_i$ and $\boldsymbol{v}_i$ are independent, the expected value of the cross products between $\boldsymbol{e}_{i-1}$ and both $\boldsymbol{\eta}_i$ and $\boldsymbol{v}_i$ vanishes. Then, multiplying (10) by its transposed, the covariance matrix of the error is

$$\Phi_i = \begin{pmatrix} I & -K_i \end{pmatrix} \left[\begin{pmatrix} I \\ C \end{pmatrix} \Phi_{i-1} \begin{pmatrix} I \\ C \end{pmatrix}^T + \begin{pmatrix} Q_i & Q_i C^T \\ CQ_i & CQ_i C^T + R_i \end{pmatrix} \right] \begin{pmatrix} I \\ -K_i^T \end{pmatrix}, \tag{11}$$

where we use the definition of the noise covariance matrices, $Q_i = \mathcal{E}[\boldsymbol{\eta}_i \boldsymbol{\eta}_i^T]$ and $R_i = \mathcal{E}[\boldsymbol{v}_i \boldsymbol{v}_i^T]$. From (11) it follows that if Φ_{i-1} is positive definite, then Φ_i is also positive definite.
Given expression (11), we want to find K_i such that Φ_i is minimized. This is a convex quadratic minimization problem with unique solution given in Åström and Wittenmark 1997 page 430. The solution is,

$$K_i = (\Phi_{i-1} + Q_i)C^T(C(\Phi_{i-1} + Q_i)C^T + R_i)^{-1}, \tag{12}$$

$$\hat{\boldsymbol{x}}_i = \hat{\boldsymbol{x}}_{i-1} + K_i(\tilde{\boldsymbol{y}}_i - C\hat{\boldsymbol{x}}_{i-1}), \tag{13}$$

$$\Phi_i = (I - K_i C)(\Phi_{i-1} + Q_i). \tag{14}$$

The equations above are applied in each generation to obtain the best predictions possible.

Appendix B

As explained in the main text, the matrices are assumed to be diagonal with equal elements in the diagonal, that is $Q_i = q_i\, I$ and $R_i = r_i\, I$, where I is the 2×2 identity matrix. If we use these definitions we can rewrite the equations from Appendix A as,

$$K_i = (\Phi^*_{i-1} + \frac{q_i}{r_i} I)C^T(C(\Phi^*_{i-1} + \frac{q_i}{r_i} I)C^T + I)^{-1}, \tag{15}$$

$$\hat{\boldsymbol{x}}_i = \hat{\boldsymbol{x}}_{i-1} + K_i(\tilde{\boldsymbol{y}}_i - C\hat{\boldsymbol{x}}_{i-1}), \tag{16}$$

$$\Phi^*_i = (I - K_i C)(\Phi^*_{i-1} + \frac{q_i}{r_i} I). \tag{17}$$

Where we define $\Phi^*_i = \Phi_i / r_i$. Written in this form, the optimization problem is reduced to a single variable, namely the quotient q_i/r_i. That is, in the window of past generations we try several values of q_i/r_i and keep the value that results in the smallest prediction error. The process is repeated for each i of a given time series.

References

Aguirre, J. D., Hine, E., McGuigan, K., & Blows, M. W. (2013). Comparing G: multivariate analysis of genetic variation in multiple populations. Heredity 2014 112:1, 112(1), 21–29.

Assis, A. P. A., Patton, J. L., Hubbe, A., & Marroig, G. (2016). Directional selection effects on patterns of phenotypic (co)variation in wild populations. Proceedings of the Royal Society B: Biological Sciences, 283(1843).

Åström, K.,& Wittenmark, B. (1997). Computer-controlled systems: theory and design. 3rd Edition. Englewood Cliffs, NK: Prentice-Hall.

Bonnet, T., Morrissey, M. B., Morris, A., Morris, S., Clutton-Brock, T. H., Pemberton, J. M., & Kruuk, L. E. B. (2019). The role of selection and evolution in changing parturition date in a red deer population. PLoS Biology, 17(11).

Cheverud, J. M. (1988). A comparison of genetic and phenotypic correlations. Evolution, 42(5), 958–968. Cobey, S. (2020). Modeling infectious disease dynamics. Science, 368(6492), 713-714.

Eisen, E. J. (1972). Long-Term Selection Response for 12-Day Litter Weight in Mice. Genetics, 72(1), 129. Ghysels, E., & Marcellino, M. (2018). Applied economic forecasting using time series methods. Oxford University Press.

Gimelfarb, A., & Willis, J. H. (1994). Linearity Versus nonlinearity of offspring-parent regression: an experimental study of Drosophila melanogaster. Genetics, 138(2), 343-352.

Gomulkiewicz, R., & Shaw, R. G. (2013). Evolutionary rescue beyond the models. Philosophical Transactions of the Royal Society B: Biological Sciences, 368(1610).

Grassini, P., Eskridge, K. M., & Cassman, K. G. (2013). Distinguishing between yield advances and yield plateaus in historical crop production trends. Nature communications, 4(1), 1-11.

Grewal, M. S., & Andrews, A. P. (2010). Applications of Kalman Filtering in Aerospace 1960 to the Present. IEEE Control Systems, 30(3), 69–78.

Harjunmaa, E., Seidel, K., Häkkinen, T., Renvoisé, E., Corfe, I. J., Kallonen, A., ... & Jernvall, J. (2014). Replaying evolutionary transitions from the dental fossil record. Nature, 512(7512), 44-48.

Hayati, M., Biller, P., & Colijn, C. (2020). Predicting the short-term success of human influenza virus variants with machine learning. Proceedings of the Royal Society B, 287(1924).

Heywood, J. S. (2005). An exact form of the breeder's equation for the evolution of a quantitative trait under natural selection. Evolution, 59(11), 2287-2298.

Houle, D., & Meyer, K. (2015). Estimating sampling error of evolutionary statistics based on genetic covariance matrices using maximum likelihood. Journal of Evolutionary Biology, 28(8), 1542-1549.

Houle, D., Mezey, J., Galpern, P., & Carter, A. (2003). Automated measurement of Drosophila wings. BMC Evolutionary Biology, 3.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. Journal of Fluids Engineering, Transactions of the ASME, 82(1), 35–45.

Kruuk, L. E. B. (2004). Estimating genetic parameters in natural populations using the "animal model". Philosophical Transactions of the Royal Society B: Biological Sciences, 359(1446), 873.

Lande, R. (1979). Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. Evolution, 402-416.

Lande, R., & Arnold, S. J. (1983). The measurement of selection on correlated characters. Evolution, 1210-1226.

Lässig, M., Mustonen, V., & Walczak, A. M. (2017). Predicting evolution. Nature ecology & evolution, 1(3), 1-9.

Le Rouzic, A., Houle, D., & Hansen, T. F. (2011). A modelling framework for the analysis of artificial-selection time series. Genetics research, 93(2), 155-173.

Le Rouzic, A., Renneville, C., Millot, A., Agostini, S., Carmignac, D., & Édeline, É. (2020). Unidirectional response to bidirectional selection on body size II. Quantitative genetics. Ecology and Evolution, 10(20), 11453–11466.

Love, A. C., Grabowski, M., Houle, D., Liow, L. H., Porto, A., Tsuboi, M., Voje, K. L., & Hunt, G. (2021). Evolvability and the Fossil Record. Preprint in EcoEvoRxiv.

Lush, J. L. (1937). Animal breeding plans. In Animal breeding plans. (Issue Edn 2). Iowa State College Press.

Lynch, M., & Walsh, B. (1998). Genetics and analysis of quantitative traits. Oxford University Press.

Merilä, J., Sheldon, B. C., & Kruuk, L. E. B. (2001). Explaining stasis: microevolutionary studies in natural populations. Genetica, 112, 199–222.

Meyer, K. (2007). WOMBAT—A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). Journal of Zhejiang University. Science. B, 8(11), 815.

Mezey, J. G., & Houle, D. (2005). The dimensionality of genetic variation for wing shape in Drosophila melanogaster. Evolution, 59(5), 1027–1038.

Milocco, L., & Salazar-Ciudad, I. (2020). Is evolution predictable? Quantitative genetics under complex genotype-phenotype maps. Evolution, 74(2), 230-244.

Milocco, L., & Salazar-Ciudad, I. (2021). The evolution of the G-matrix under nonlinear genotype-phenotype maps. The American Naturalist. In press.

Morrissey, M. B., Kruuk, L. E. B., & Wilson, A. J. (2010). The danger of applying the breeder's equation in observational studies of natural populations. Journal of Evolutionary Biology, 23(11), 2277–2288.

Nosil, P., Villoutreix, R., Carvalho, C. F. de, Farkas, T. E., Soria-Carrasco, V., Feder, J. L., Crespi, B. J., & Gompert, Z. (2018). Natural selection and the predictability of evolution in Timema stick insects. Science, 359(6377), 765–770.

Nosil, P., Villoutreix, R., de Carvalho, C. F., Feder, J. L., Parchman, T. L., & Gompert, Z. (2020). Ecology shapes epistasis in a genotype–phenotype–fitness map for stick insect colour. Nature Ecology & Evolution, 4(12), 1673-1684.

Pélabon, C., Albertsen, E., Rouzic, A. Le, Firmat, C., Bolstad, G. H., Armbruster, W. S., & Hansen, T. F. (2021). Quantitative assessment of observed versus predicted responses to selection. Evolution, 75(9), 2217–2236.

Pigliucci, M. (2006). Genetic Variance–covariance Matrices: A Critique of the Evolutionary Quantitative Genetics Research Program. Biology and Philosophy 2006 21:1, 21(1), 1–23.

Pujol, B., Blanchet, S., Charmantier, A., Danchin, E., Facon, B., Marrot, P., Roux, F., Scotti, I., Teplitsky, C., Thomson, C. E., & Winney, I. (2018). The Missing Response to Selection in the Wild. Trends in Ecology & Evolution, 33(5), 337–346.

Rescan, M., Grulois, D., Ortega-Aboud, E., & Chevin, L. M. (2020). Phenotypic memory drives population growth and extinction risk in a noisy environment. Nature ecology & evolution, 4(2), 193-201.

Rescan, M., Grulois, D., Aboud, E. O., de Villemereuil, P., & Chevin, L. M. (2021). Predicting population genetic change in an autocorrelated random environment: Insights from a large automated experiment. PLoS genetics, 17(6), e1009611.

Rice, S. H. (2004). Evolutionary Theory: Mathematical and Conceptual Foundations. Sinauer Associates.

Rice, S. H. (2012). The place of development in mathematical evolutionary theory. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution, 318(6), 480-488.

Roff, D. A. (2007). A centennial celebration for quantitative genetics. Evolution, 61(5), 1017–1032.

Rutledge, J. J., Eisen, E. J., & Legates, J. E. (1973). An experimental evaluation of genetic correlation. Genetics, 75(4), 709-726.

Salazar-Ciudad, I., & Jernvall, J. (2010). A computational model of teeth and the developmental origins of morphological variation. Nature 2010 464:7288, 464(7288), 583–586.

Sgrò, C. M., & Hoffmann, A. A. (2004). Genetic correlations, tradeoffs and environmental variation. Heredity, 93(3), 241-248.

Shaw, R. G. (2019). From the past to the future: Considering the value and limits of evolutionary prediction. American Naturalist, 193(1), 1–10.

Sodini, S. M., Kemper, K. E., Wray, N. R., Trzaskowski, M. (2018). Comparison of genotypic and phenotypic correlations: Cheverud's conjecture in humans. Genetics, 209(3), 941-948.

Steppan, S. J., Phillips, P. C., & Houle, D. (2002). Comparative quantitative genetics: evolution of the G matrix. Trends in Ecology & Evolution, 17(7), 320-327.

Walsh, B., & Lynch, M. (2018). Evolution and selection of quantitative traits. Oxford University Press.

Wood, C. W., & Brodie, E. D. (2015). Environmental effects on the structure of the G-matrix. Evolution, 69(11), 2927–2940.

Wortel, M., Agashe, D., Bailey, S., Bank, C., & Bisschop, K. (2021). The why, what and how of predicting evolution across biology: from disease to biotechnology to biodiversity. Preprint in EcoEvoRxiv.

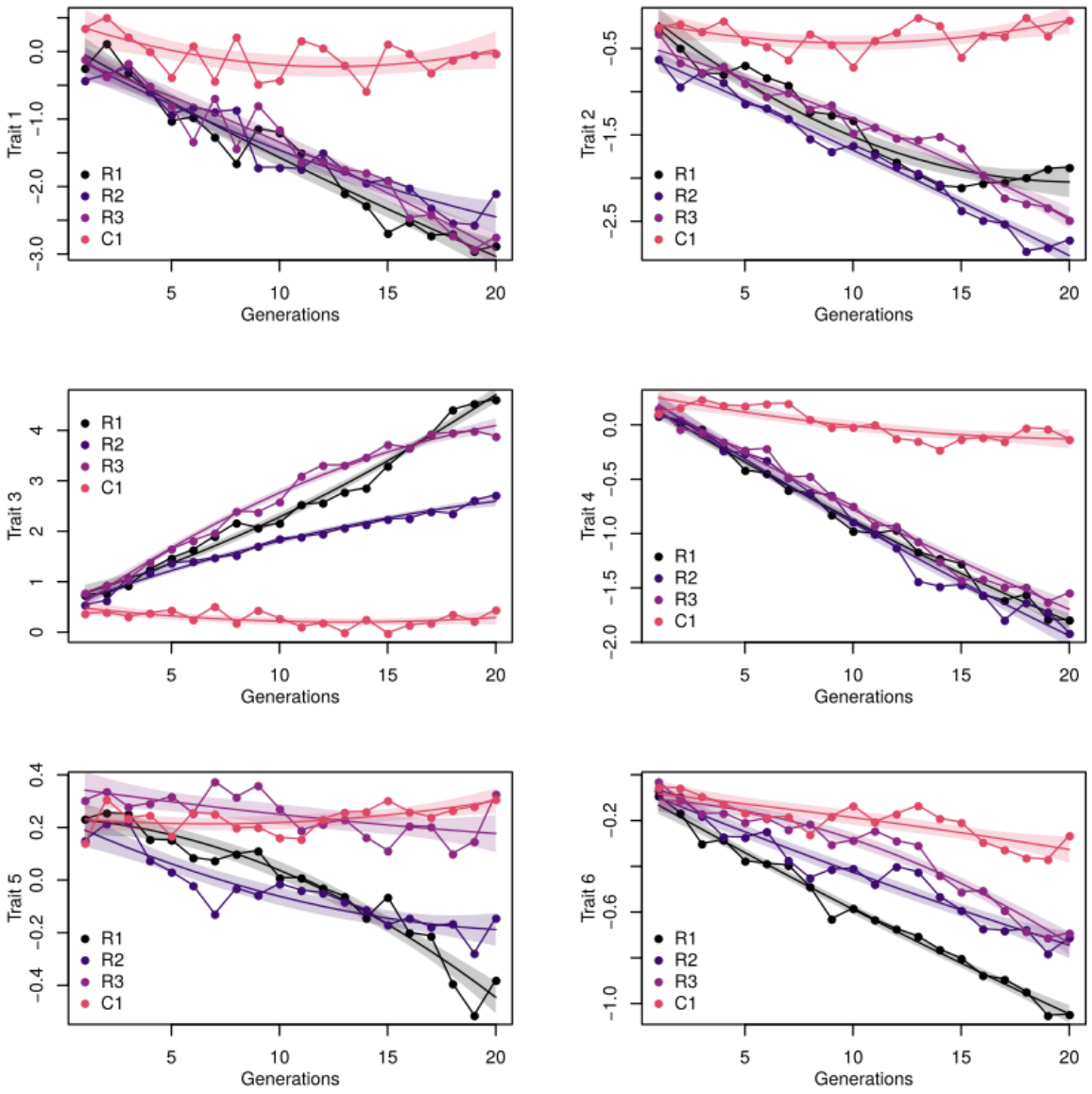
Acknowledgements

We thank Ruben H. Milocco for advice in applying the Kalman filter. We thank David Houle and his team for instructions on how to perform artificial selection experiments with the fruit fly using the WingMachine. We thank Pınar Güler Kohlmeier from the Billeter team for providing the flies from the wild population. We also thank Niko Björkell, Shuja Wahid, Sara Goncalves and Ardit Fejzullahi for help carrying out the experiments. The research was funded by the Academy of Finland, the Integrative Life Sciences graduate program of the University of Helsinki, the Spanish Ministry of Science and innovation (PGC2018-096802-B-I00), and the Spanish State Research Agency, through the Severo Ochoa and María de Maeztu Program for Centers and Units of Excellence in R&D (CEX2020-001084-M). We also thank CERCA Programme/Generalitat de Catalunya for institutional support.

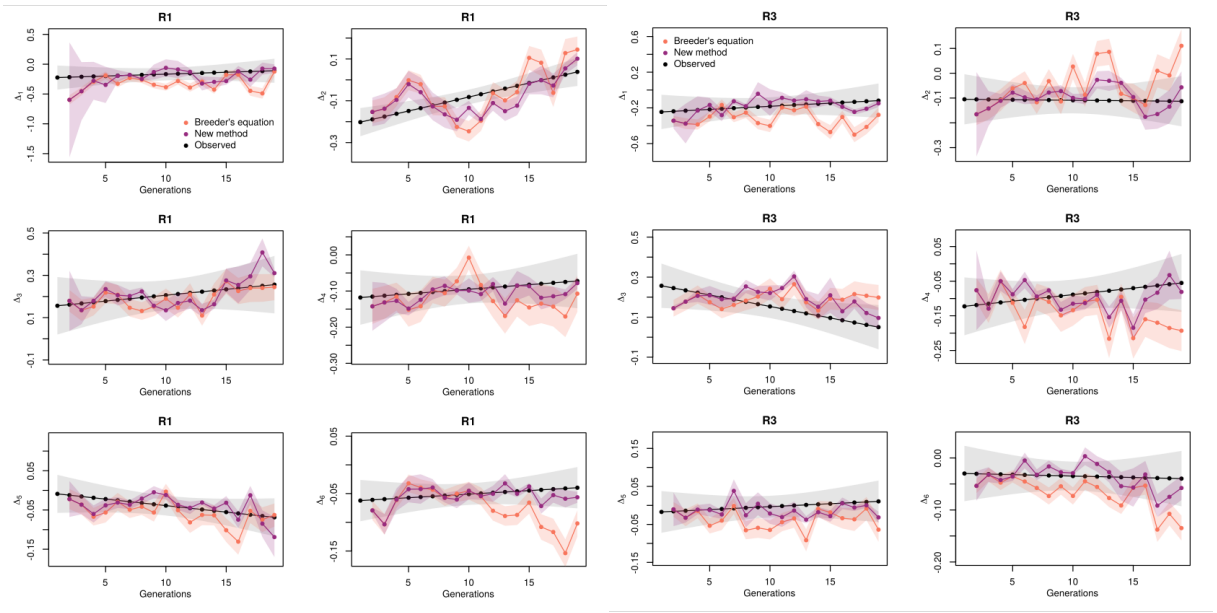
Statement of authorship

LM conceptualized and developed the method, with help from ISC. LM carried out the experiments and wrote the original draft of the manuscript. LM and ISC designed the artificial selection experiments, participated in the editing and writing of the manuscript, and obtained funding.

Supplementary Figures



Supplementary Figure 1. Change in trait mean for the 3 replicate lines (R1, R2, R3) and the control (C1) for the artificial selection experiments of the wing. The quadratic fit and its 95% CI is included.



719 **Supplementary Figure 2**. Predicted and observed responses to selection for lines R1 and R2,
720 with G estimated with a pedigree depth of 2 (i.e. like Main Figure 2). The observed change is
721 plotted with 95% CI and the predicted changes are plotted with 1 SD.