MEASURING SELECTION COEFFICIENTS BELOW 10^{-3}: METHOD, QUESTIONS AND PROSPECTS

Romain Gallet[1], Tim F. Cooper[2], Santiago F. Elena[3,4] and Thomas Lenormand[1*]

*: corresponding author

Affiliations:

[1]: CEFE - UMR 5175 1919 route de Mende, F-34293 Montpellier cedex 5, France

[2]: Evolution and Ecology Group, Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA

[3]: Instituto de Biología Molecular y Celular de Plantas (CSIC-UPV), Campus UPV CPI 8E, CL. Ingeniero Fausto Elio s/n, 46022 València, Spain

[4]: Santa Fe Institute, 1399 Hyde Road Park, Santa Fe, NM 87501, USA

Email:

rgallet@gmail.com

tfcooper@uh.edu

sfelena@ibmcp.upv.es

thomas.lenormand@cefe.cnrs.fr

Running title: High precision fitness measures

Keywords: mutation fitness effect, mixed model, high throughput data, drift, flow cytometry, transitivity, epistasis, fluorescence, *E. coli*

ABSTRACT

Measuring fitness with precision is a key issue in evolutionary biology, in particular to study mutations of small effects. It is usually thought that sampling error and drift prevent precise measurement of very small fitness effects. We circumvented these limits by using a new combined approach to measure and analyze fitness. We estimated the mutational fitness effect (MFE) of three independent mini-Tn*10* transposon insertion mutations by conducting competition experiments in large populations of *Escherichia coli* under controlled laboratory conditions. Using flow cytometry to assess genotype frequencies from very large samples alleviated the problem of sampling error, while the effect of drift was controlled by using large populations and massive replication of fitness measures. Furthermore, with a set of four competition experiments between ancestral and mutant genotypes, we were able to decompose fitness measures into four estimated parameters that account for fitness effects of our fluorescent marker (α), the mutation (β), epistasis between the mutation and the marker (γ), and departure from transitivity (τ). Our method allowed us to estimate mean selection coefficients to a precision of 2×10^{-4}. We also found small, but significant epistatic interactions between the allelic effects of mutations and markers, and confirmed that fitness effects were transitive in most cases. Unexpectedly, we also detected variation in measures of *s* that were significantly bigger than expected due to drift alone, indicating the existence of cryptic variation, even in fully controlled experiments. Overall our results indicate that selection coefficients are best understood as being distributed, representing a limit on the precision with which selection can be measured, even under controlled laboratory conditions.

INTRODUCTION

Mutations of small effect can play an important role in evolution, but they are difficult to measure experimentally because the precision with which fitness effects can be measured is relatively low (see below for details). For this reason, it remains unclear to what extent mutations with small beneficial effects contribute to fitness improvements (ORR 2005). It is also unclear how much deleterious mutations of small effect contribute to the genetic load and inbreeding depression (BATAILLON and KIRKPATRICK 2000; CHARLESWORTH and CHARLESWORTH 1998). More generally, the existence and influence of mutations of small effect is at the

heart of the neutralist-selectionist controversy (e.g. NEI 2005). This debate can only be addressed experimentally if the precision of fitness measurements is lower than the inverse of effective population size, which seems beyond reach for large populations (KREITMAN 1996). Finally, a low precision in fitness measures limits the ability to determine whether the fitness effect of a mutation varies across different environmental or genetic contexts and adds to other sources of stochasticity (LENORMAND *et al.* 2009) to make it difficult to reliably predict evolutionary trajectories.

Precisely measuring fitness poses technical, conceptual and statistical challenges. The technical challenge is to set up a technique that allows experiments to be carried out efficiently. The first major advance was to use 'population cages' with *Drosophila* or other small animals (starting in the 30's with the work of L'Heritier and Teissier (1937a; 1937b)). With such devices, environmental conditions are relatively controlled and gene flow can be eliminated. However drift and indirect selection caused by loci under selection in linkage disequilibrium with the focal locus, are difficult to account for. The same approach was applied to microorganisms (DYKHUIZEN and HARTL 1980), which can be made isogenic save for a focal gene thereby reducing indirect selection due to initial linkage disequilibrium (e.g. CARRASCO *et al.* 2007; DOMINGO-CALAP *et al.* 2009 for distribution of mutation fitness effects; ELENA *et al.* 1998; PERIS *et al.* 2010; SANJUAN *et al.* 2004) and can be propagated as large populations, minimizing the effect of drift relative to selection. They can also be followed over many generations (DYKHUIZEN and HARTL 1983; LUNZER *et al.* 2002; THATCHER *et al.* 1998). Long term monitoring increases the ability to detect small differences in fitness between competing genotypes, but adds the complication that newly arising mutations may perturb the assay (DYKHUIZEN and HARTL 1983). An important technical issue in all competition experiments is to determine the frequency of competing genotypes reliably and quickly. In many cases the idea is to link an easily recognized marker with the gene under scrutiny. It is, however, important to recognize that a marker can confer a selective difference (a marker 'cost'), which might vary with the genetic background (epistasis) or external environment (G×E interactions). Finally, inferring allelic selection coefficients against a common reference strain requires that genotypic fitness is transitive. These potential complications require adding proper controls to competition experiments.

A key conceptual difficulty in measuring the fitness effects of mutations is to distinguish selection from drift (BEATTY 1984; MILLSTEIN 2008), which is at the heart of several population cage experiments with *Drosophila* (DOBZHANSKY and PAVLOVSKY 1957). To account for the effect of drift, a selection coefficient can be defined from the expected change in allele frequency over one generation (e.g. ROUSSET 2004), which can be estimated from the mean frequency change in independent competition experiments. Because of drift, replication is fundamentally necessary to estimate fitness, and the precision of a given fitness measure must account for the inter-replicate variance. Indeed, it is possible to count all organisms in an experimental population, so that the genotype frequencies are known without sampling error. Such an experiment would allow frequency variation to be determined 'exactly', but would clearly not account for the possibility that drift will cause different outcomes in different replicates. A further complication is that fitness may vary because of changing environmental conditions. Fluctuating selection during the course of a competition experiment or varying selection across replicates of a competition assay can mimic drift (FELSENSTEIN 1976; LYNCH 1987; O'HARA 2005). If selection varies, and it probably always does to some extent (BELL 2008a; BELL 2010), measuring selection requires measuring both a mean *and* a variance (the latter not including sampling error). The remaining variance can be caused by drift or by heterogeneity in selection, which are difficult to disentangle without extra information on the effective population size. In summary, measuring selection with precision requires estimating an expectation over several replicates, so that its variance can be decomposed into components due to sampling error, drift and variable selection.

From a statistical point of view, selection coefficients in the field or in the laboratory are best estimated by using a fully specified selection model in a likelihood framework (e.g.ARNASON and LEWONTIN 1991; CLARK 1979; LABBE *et al.* 2009; LENORMAND and RAYMOND 2000; MANLY 1985; OAKESHOTT *et al.* 1983; SACCHERI *et al.* 2008; WILSON *et al.* 1982), which can include drift if longitudinal data are available (BOLLBACK *et al.* 2008; MANLY 1985; O'HARA 2005). When selection can be approximated by a continuous process through time in an isolated population, a simple approach is to regress Log(p/q) (where p and q represent the frequencies of the two competitors) over time expressed in units of generations (FISHER 1930). The connection with logistic regression and general linear models is then straightforward (ARNASON and BARKER 1999) and more appropriate than the use of least

squares. However, complications arise in the analysis of time series and correlated error in repeated measurement through time (ARNASON and BARKER 1999; O'HARA 2005), especially when both drift and fluctuating selection cause frequency variation. The latter problems can be important, particularly when analyzing multiple time point series (e.g., arising in long term population cage or chemostat experiments), although they are rarely taken into account. Often, replicated experiments are simply pooled, even if significantly different, and not analyzed to consider variance in the estimates of selection. The development of mixed models offers an attractive alternative to circumvent this problem and to measure selection and its variation.

We present an approach combining several features to improve and quantify the precision of fitness measures. First, we use techniques that have proved to be among the most efficient to measure fitness: competition assay between large populations of *Escherichia coli* strains to minimize drift and engineered mutations to avoid the problem of indirect selection. Specifically, we used three genotypes each carrying a single mutation introduced by the integration of a mini-Tn*10* transposon. These mutations were considered neutral, relative to a common progenitor genotype, in a previous experiment (ELENA *et al.* 1998). We use two fluorescent markers (ROSENFELD *et al.* 2005) combined with flow cytometry (LUNZER *et al.* 2002) to measure frequency variation with great precision, and thus minimize sampling error. Other studies have shown the utility of these approaches in measuring genotype fitness (LEE *et al.* 2009; LUNZER *et al.* 2002; ZHU *et al.* 2005). Key aspects of our approach are (1) a comprehensive set of four competition assays that enable us to separately estimate mutational selection coefficients (α), the cost of the marker (β), epistasis between mutation and marker (γ), and transitivity (τ). (2) We use short-term batch culture to facilitate massive replication and to reduce the possibility that *de novo* beneficial mutations will occur. (3) We analyze the data in an integrated likelihood framework with random effects to partition sources of variation in our estimates (sampling error versus drift versus variable selection). Our approach allowed us to estimate both mean *and* variance in selection coefficients at a precision of 0.02%. This precision allowed us to detect variation in measures of some mutation selection coefficients that were significantly larger than expected due to drift alone, indicating the action of some kind of cryptic variation during our competitions. This finding implies that, in practice, selection coefficients should be considered as being

distributed, and that precise measures requires evaluating both the mean *and* the variance of this distribution. Furthermore, the variance in *s* indicates that some uncontrolled processes occur in these experiments (cryptic environmental or genetic variation), which impose a limit to further dissecting the differences seen across replicates. We discuss implications of these findings and the prospects of this high-throughput method for fitness measurement.

MATERIAL AND METHODS

Strain construction

The *E. coli* B strain used in the present study, REL4548, was evolved in Davis Minimal medium supplemented with 25 μg/mL glucose (DM25) for 10 000 generations as part of a long-term evolution experiment (ELENA *et al.* 1998).

Insertions of the chromosomal fluorescent markers : The YFP and CFP genes (provided by the Yeast Resource Center of the University of Washington) were inserted at the *rhaA* locus of REL4548 using the a technique developed by Datsenko and Wanner (DATSENKO and WANNER 2000). Table 1 gives a description of this method as applied to our experiments. A full description of the method is given in *supplementary material*.

Mutant construction: The three mutants studied here were constructed by Elena *et al.* (ELENA *et al.* 1998) and were obtained by random single insertions of mini-Tn*10* derivative 104 – which contains a tetracycline resistance cassette (KLECKNER *et al.* 1991) – into REL4548. We chose mutations T63, T103, and T121 from this original collection because they were identified as neutral using the standard plating method. These mutations were transduced into REL4548/CFP and REL4548/YFP by P1 transduction, in order to have each mutation associated with each fluorescent marker. Since P1 transductions were performed between isogenic strains (except for the marker and the mobilized mutation), the risk of secondary mutations was low. Transductants were selected on LBA-Tet plates (LB agar plates supplemented 10 μg/mL Tetracycline). We denote '*wc*' the wild type genotype with CFP marker (*wc* for wild type cyan), '*wy*' for wild type YFP, '*mc*' for mutant CFP, '*my*' for mutant YFP.

Competition experiments

Media: Lysogeny broth (LB) was used for routine molecular work and for reviving strains from storage (10 g/L NaCl, 10 g/L tryptone, 5 g/L yeast extract; LB Agar LB + 15 g/L agar). Davis minimal (DM) medium supplemented with 250 μg/mL glucose (DM 250) was used for all competition assays ($KH_2PO_4 \cdot 3H_20$ 7 g/L, KH_2PO_4 2 g/L, $(NH_4)_2SO_4$ 1 g/L, sodium citrate 0.5 g/L; pH was adjusted to 7.0 with HCl or NaOH as necessary). Bottles were weighed before and after autoclaving and sterile milliQ water was added to compensate for evaporation. After autoclaving, DM was supplemented with: 2.5 mL glucose 10%, 1 mL $MgSO_4^{2-}$ 10%, 1 mL Thiamine (vitamin B1) 0.2%. We call this medium DM250, which is equivalent to the one used by Lenski (1991), in which the strain REL4548 grew for 10 000 generations, but with 10 times more glucose. Glycerol stocks: All strains were grown overnight and a sample of 750 μL of each culture was mixed to 250 μL of 60% glycerol and kept at -80°C for storage.

Culture: The relative fitness, *W*, of each mutant was estimated by measuring the change in its relative frequency in competition experiments. In order to measure the mutation fitness effect (MFE) and to control for potential marker effects and epistasis between the mutation and the marker, we performed four competitions types for each mutant: (a) *wc*/*wy*, (b) *mc*/*my*, (c) *my/wc*, and (d) *mc*/*wy*. The rationale for performing all these competitions is presented below. Competitions were begun by growing the strains to be competed at 37°C overnight with shaking at 250 rpm in 24 well microtiter plates (Greiner Bio-one 662102 - suspension culture plates) containing 1 mL/well of DM250. We used DM250 as the growth medium to obtain large population sizes, which limit the effect of drift, and to facilitate the measurement of hundreds of thousands cells without having to sample large volumes. To limit evaporation, each 24-well plate was placed in a two-liter plastic box containing paper towels soaked with 100 mL water (at the bottom of the box). The next day, 10 μL (100-fold dilution) of each culture was transferred to a fresh plate and incubated for 24 hours under identical conditions. On the third day, competitors were mixed at a 1:1 ratio (5 μL of each competitor) and transferred to a fresh plate under identical conditions. On day four, 20 μL of each competition was transferred into 10 replicate wells containing 1980 μL of DM250. After mixing, 1 mL was removed from each well and placed in a plastic test tube at 4°C for a subsequent flow cytometry measurement (performed one hour later), while the remaining 1 mL was kept in the microtiter plate to be cultivated under the conditions described above.

Finally, on the fifth day, a 100 μL sample was taken from each competition, diluted in DM (not containing glucose, thiamine or $MgSO_4^{2-}$), and placed in a plastic test tube at 4°C for a subsequent flow cytometry measurement (performed one hour later). Ten different types of competitions were performed: *wc* vs *wy*, $m_{T63}c$ vs $m_{T63}y$, $m_{T63}c$ vs *wy*, *wc* vs $m_{T63}y$, $m_{T103}c$ vs $m_{T103}y$, $m_{T103}c$ vs *wy*, *wc* vs $m_{T103}y$, $m_{T121}c$ vs $m_{T121}y$, $m_{T121}c$ vs $m_{T121}y$, *wc* vs $m_{T121}y$. Each experimental block consisted of each of these 10 competitions replicated 10-fold. Each experimental block was repeated at four different dates.

Flow cytometry: The relative frequency of competitors marked with CFP or YFP was measured using a Gallios Beckman Coulter flow cytometer at 0 and 24 hours following mixing of competing genotypes. We decided to separate competitor populations only on the basis of their fluorescent markers, because CFP and YFP cell populations did not have the same distribution pattern on forward (FCS) vs. side scatter (SSC) plots. Thresholds were applied manually (since clustering algorithms often introduce more noise) on the CFP – YFP plots to determine the boundaries of each population (CFP, YFP, unmarked cells and doubled marked objects) as shown on Figure 1. These thresholds were the same for all competition plots because in such a constant environment, cell clusters were always localized in the same areas of the plot. The frequency of each marker type was calculated using CFP and YFP population counts only. Unmarked and double-marked populations represented approximately 0.2 and 1% of the total population, respectively. For simplicity, 'doublets' (objects composed of two cells) were excluded from our frequency estimates. CC, YY and CY doublets occur, but only the latter are detected in the C2 population (Figure 1). Furthermore, doublets may not form at random; doublets with the same color were often overrepresented (data not shown). Nevertheless, even considering these complications, ignoring doublets only introduces a bias on s measures proportional to $s\varepsilon$, where ε is the fraction of the CY population (C2 in Figure 1). Under our conditions, $\varepsilon \approx 1\%$ making this bias negligible compared to s (see supplementary materials for details).

Precision of frequency measures with cytometry

Our method is based on measuring the relative frequency p of two competing genotypes at different time points by counting $C = 200000$ cells. This large figure, however, still represents

a small fraction of the total population and therefore, we estimate frequencies with sampling error. The theoretical expectation for this sampling error is

$$\sigma_e^2 = p(1-p)/C.$$

If nothing else contributes to measurement error, we should obtain this variance when measuring repeatedly the frequency in a given test tube. Preliminary experiments (not shown) indicated that much larger error could occur, in particular when test tubes were insufficiently mixed. This is an important technical issue and comparing actual measurement error to σ_e^2 provides an internal check that measurement error is not inflated above the sampling error expectation. In the experiments presented here, we used measures of initial frequencies (p_0) in our replicated competitions to estimate the variance of frequency measures performed with cytometry σ_{obs}^2. We found that $\sigma_{obs}^2/\sigma_e^2$ was 0.94, 1.83, 1.07, and 0.95 for the four different dates where all the competitions were performed. Except for date 2, measurement error was very close to that inherent to sampling only. However as shown by σ_{obs}^2 at date 2 (and other preliminary assays showing more dramatic results), using cytometry does not guarantee that measurement error will be low. In particular, thorough mixing of test tubes throughout the growth cycle limits cell aggregation and is a crucial step in taking advantage of the advantages offered by the cytometric approach (or any other approach based on frequency variation).

Measure of genotypic fitness

We measured fitness based on a continuous time model $dp/dt = s\,p\,(1-p)$, which defines selection coefficient (s) based on frequency (p) variation. This frequency variation was measured in the competition experiments described above. For a given competition assay k, the data is a vector $\mathbf{n}_k = \{n_{1k}^0, n_{2k}^0, n_{1k}^t, n_{2k}^t\}$ giving the number of genotypes 1 and 2 counted at time 0 (beginning) and t (end of the competition). The log-likelihood of this data given initial frequency of genotype 1 p_{1k}^0 and selection coefficient s_k is computed as

$$\ln \Pr(\mathbf{n}_k | p_{1k}^0, s_k) = \sum_{i=1,2} \sum_{j=0,t} \left(n_{ik}^j \ln p_{ik}^j \right) , \quad (1)$$

where $p_{2k}^j = 1 - p_{1k}^j$ and

$$p_{1k}^{t} = \mathrm{e}^{s_k t} p_{1k}^{0} / (p_{2k}^{0} + \mathrm{e}^{s_k t} p_{1k}^{0}) \,. \qquad (2)$$

The frequency variation is measured over 24 h. In order to scale fitness measurements 'per generation', we used the number of cell generations as the time unit. This measure is an average over the duration of the competition, which does not contradict the fact that conditions change with time in a given assay (e.g. the glucose becomes limiting) because it does so similarly in all replicates. Because populations expand by binary fission, we have t = ln(100)/ln(2) = 6.6 in eq. (2). Across replicates of the same competition, s_k might vary for reasons other than sampling error, owing, for example, to drift or to cryptic environmental variation. To measure this variation, we used the same logistic regression approach (eqn 1-2), but including the assumption that s was normally distributed $s \sim N(\overline{s}, \sigma_s)$ among replicates. The log-likelihood of this logistic regression with random slope is then

$$\ln \Pr(\mathbf{n} | \mathbf{p}_1^0, \overline{s}, \sigma_s) = \sum_k \ln \int_{-\infty}^{\infty} N(\overline{s}, \sigma_s; s) \prod_{i=1,2} \prod_{j=0,t} n_{ik}^{j} \ln p_{ik}^{j} \, ds \,, \qquad (3)$$

where $\mathbf{n}$ is the data matrix $\{\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3, \ldots\}$, $\mathbf{p}_1^0$ the vector of all initial frequencies and $N(\mu, \sigma; x)$ denotes the probability density function of the Normal distribution with mean μ and standard deviation σ. In all cases, parameters were estimated by maximizing the log-likelihood. Support limits for a given estimate were computed within 2 units of log-likelihood from the maximum with all other parameters being freely fitted. An equivalent of 'standard error' SE_{eq} was computed as a quarter of the support range (similarly, 95% confidence intervals are ±1.96 SE). Computations were done using Mathematica (WOLFRAM RESEARCH 2008).

Fitness transitivity, allelic fitness and epistasis

To test whether a constant fitness can be attributed to a genotype irrespective of its competitor, we performed all possible combinations of competition assays for a given

mutant. At a first locus we have the wild type (*w*) and mutant (*m*) alleles. At a second locus we have two alleles *c* and *y* (corresponding to the CFP and YFP marker proteins, respectively). Each competition assay requires competing genotypes to have different alleles at the marker locus. There are thus four possible combinations: (a) *wc*/*wy*, (b) *mc*/*my*, (c) *my/wc*, and (d) *mc*/*wy*. Table 2 indicates the selection coefficient expected in each of these cases if we assume that the fitness of genotypes *wc*, *wy*, *mc* and *my* are constant and equal to W_{wc}, W_{wy}, W_{mc}, and W_{my}, respectively. When measuring the marker effect in the same background (competitions (a) and (b)), we measured the selection coefficient of the CFP genotype. Otherwise, we measured the selection coefficient of the mutant genotype against the wild type (competitions (c) and (d)).

Population genetics models usually assume that fitness effects are transitive, *i.e.* that they could be deduced from some absolute value ranking of the different genotypes. However this is an assumption that requires evaluation before attributing a selection coefficient to genotypes. Since competitions (a), (b) and (c) are sufficient to estimate all fitness if they are transitive, competition (d) can be used to measure departure from transitivity. Specifically, we introduce a parameter τ measuring this departure (see table 2). Further re-parameterization allows decomposing genotypic fitness into allelic effects and their interaction (epistasis). We note $W_{wc} = W_{wy} + \alpha$, $W_{my} = W_{wy} + \beta$, $W_{mc} = W_{wy} + \alpha + \beta + \gamma$. α is the "cost" of the CFP marker, β is the selective effect of the mini-Tn*10* mutation and γ is the epistasis between the two loci.

To fit this model, for each mutant, we used eq. (1) summed over the four competition assays and their replicates, with the parameterization indicated above. Support limits for estimates were computed within 2 units of log-likelihood all other parameters being freely fitted.

Expected amount of drift

In our experiments, population size increases by binary fission. To compute the variance in frequency introduced by drift, we first determine that each bacteria division increases this variance by a quantity pq/n^2, where *n* is the population size at the time of this division. We then sum this variance to the end of population growth:

$$\sum_{n=n_i}^{n_f} \frac{pq}{n^2} = pq\left(\frac{1}{n_i} - \frac{1}{n_f}\right) + 0\left(\frac{1}{n_i}\right)^2 \quad . \tag{4}$$

We thus expect the variance of selection coefficients contributed by drift to be

$$\sigma_s^2 = \frac{n_f - n_i}{g^2 n_f n_i pq} \ , \tag{5}$$

where *g* is the number of 'generations' (6.6 as explained above over the time course of the competition experiment). The variance in frequency change caused by selective differences among replicates is var(*sgpq*) = $\sigma_s^2(gpq)^2$. In our experiments we have n_f in the range $10^8 - 10^9$ and n_i 100 times less. These population sizes were estimated by serial dilution and plating (not shown), and these numbers represent the extreme cases. Thus we expect σ_s to be between 10^{-4} and 3×10^{-4}. Significantly larger σ_s would indicate that a source of variation, in addition to sampling error and drift, contributed to differences among replicate competitions (*e.g.* such as random fluctuations in selection coefficients among replicates).

RESULTS

We used a flow cytometric approach to measure the fitness of three mutants, each carrying a single mutation, that were classified as being 'neutral' with conventional methods (ELENA *et al.* 1998). We did 10 types of competition assays, each replicated 10-fold at each of four weeks, giving a total of 400 fitness measures (Figure 2). A standard analysis of deviance (eq. 1-2) revealed that 96.6% of the deviance was among competition assay types. There were significant week (0.6% of the total deviance), week × competition (1.3%) and replicate (1.4%) effects, although they accounted for a very small fraction of the total deviance. In particular, although detectable, the week effect was smaller than the replicate effect (well-to-well variation for the same competition during the same week), indicating that the experiments were repeatable from one week to another. We also used a simple one-way ANOVA to test whether the standard deviation in s measures among replicates was consistent when measured at different weeks. This was the case, although the repeatability was not

extremely high. Specifically, we found that 53% of the variance in this standard deviation was among competitions and 47% within competition between weeks. This variation among competitions is significantly larger than across dates for the same competition ($F_{9,30}$ = 3.7, *P* = 0.003). Repeatability of means and variance at different weeks is crucial for measuring fitness with precision: it is fairly easy to obtain a very precise measure of frequency change in a single assay (or even an exact measure if all individuals in the competition are counted at the beginning and the end), but this is not equivalent to obtaining an accurate measure of fitness, which must account for inter-replicate variance. As this example shows, analyzing a very large dataset also provides sufficient statistical power to detect very small biological effects, but it can also reveal 'nuisance' effects (almost anything tested becoming 'significant'). To cope with these issues, we used an approach quantifying variance components in a mixed model (Eq. 3 in methods).

Precision of fitness measures: Competition assays provide a direct measure of the fitness of one genotype relative to a competing genotype. To determine if the differences we observed in fitness estimates between replicate competitions was biologically meaningful, as opposed to a sampling effect, we used a mixed model to directly estimate the amount of variation in *s* (σ_s) beyond sampling error (eq. 3). This approach provides an estimate of average selection intensity ($\overline{s}$), a measure of biological heterogeneity in selection among replicates (σ_s) and standard errors associated with these two parameters. Beyond estimating average selection ($\overline{s}$) with some precision, it is important to indicate the magnitude of variation in *s* (σ_s) and the precision reached to estimate it. Table 3 presents these estimates for our 10 competition assays. Estimates of $\overline{s}$ range from 0.00088 (T121 YFP) to -0.024 (T103 CFP). Estimates of σ_s range from 0 to 0.0035, with seven of ten estimates being greater than zero. In all cases the precision of these estimates is about ± 0.0002.

The origin of variation in *s* among replicates: There are four non-exclusive reasons that changes in the frequency of reference and mutant genotypes during competitions could be 'truly' different among replicates: (1) experimental error unrelated to sampling (e.g., pipeting), (2) new mutations occurring in some replicated competitions, (3) drift, and (4) variation in selection intensity among replicates (e.g., due to cryptic environmental variations). We consider each possibility in turn.

Experimental error is unlikely to be the source of the variation in s in our experiment. We repeated each competition type at four dates and σ_s was consistently low and comparable to the drift expectation in competitions (a) and (b) (Fig. 2). It is unlikely that systematic error would only occur for some competition types and even more unlikely that this pattern would be repeatable at different dates. The second hypothesis is that new deleterious or beneficial mutations, unrelated to the mutation of interest, may occur during the competition and influence the outcome. The case of deleterious mutations is not really problematic, because they are unlikely to reach high frequency in a large population and because, if many occur, they will occur equally in the two competing genotypes. The case of beneficial mutations may, at first sight, seem trickier. Let us consider a worst-case scenario of the early occurrence of a beneficial mutation providing a growth advantage of 10% per division. If we consider the appearance of this mutant at the very start of the pre-culture (*i.e.*, ~17 generations before the start of the competition assay), its frequency at the end of the competition will be $< 10^{-5}$ (assuming a competition of 6.6 generations and an effective population size of 10^6 as used in this study), which is too low to have any impact on our measures. Significant frequency variation (above ~0.02% in our case) would require a mutation to confer a benefit greater than ~30% (see supplementary materials for details), which is very unlikely in a strain that has been adapted to the environment for 10,000 generations and for which no such mutations have been identified during the early stages of this adaptation, when fitness increases were most rapid (BARRICK *et al.* 2009). Moreover, even if such large effect mutations were available to our strains, they would have to occur repeatedly in many competitions because our observed var(*s*) is not due to isolated outliers (Figure 2). We note that some mutation types, notably genomic amplifications, have been observed to occur at high frequencies and may sometimes confer beneficial effects either directly or indirectly by increasing the mutational target for new mutations to occur. However, if these mutations occur at a very high rate, they would occur in both competitors and thus have a limited effect on var(*s*). Furthermore, if the occurrence of *de novo* genomic amplifications were increasing var(*s*) in our experiments, they should do so in all competitions types, and not only in competitions (c) and (d) (Figure 3). In summary, we conclude that the rise and spread of new mutations is very unlikely to explain our results.

Drift can also cause variation in genotype frequency changes in the different replicates. This process scales with the inverse of population size and should effectively vanish in very large populations. In our experiments, we expect σ_s to be between 10^{-4} and 3×10^{-4} if it was due to drift alone (see material and methods). Our estimates of σ_s (Table 3) varied among the competition assays. In competition (a) and (b), estimates of σ_s were not different from the maximum value that would be expected because of drift (3×10^{-4}, Figure 3). These competition assays correspond to CFP versus YFP competitions within the same genetic background. We thus conclude that the cost of expressing the different fluorescent proteins is not significantly affected by uncontrolled cryptic environmental variation in our experiments. Other estimates of σ_s (in competitions (c) and (d)) are much larger than the drift expectation (Figure 3). One possibility is that the effect of drift is greater than expected from consideration of population size alone. This may be the case if there was substantial phenotypic diversity in the competing populations so that a subset of the population contributed disproportionately to population growth. In fact, this explanation seems unlikely. We find a typical value of σ_s of about 0.001, which would require that N_e was reduced to ~9% of the actual population (from eq. 5) (Figure 3). This means that drift can only explain our observed σ_s if more than 90% of the sampled population is not growing. (In the most extreme case, less than 1% of the population would have to be growing (T63, competition c)). Studies performed on *E. coli* populations showing that only a few percent of the total population were in an "atypical" non-growing physiological state during exponential population growth (BALABAN *et al.* 2004; LEVIN and ROZEN 2006), support the conclusion that phenotypic variation is not sufficient to account for variance among replicates in some of our competitions.

In cases where variation is too high to be explained by drift (competitions (c) and (d)), variation necessarily implies that selection intensity changes slightly among replicates, perhaps due to environmental variation among replicates. Furthermore, σ_s estimates were larger for large $|\bar{s}|$ (Pearson r = 0.69), a situation that would be expected when different competitors have environmental tolerance curves with different slopes (*i.e.* a GxE effect). In this case environmental variation will not necessary impact both competitors with the same intensity. Thus, small environmental variations across replicated competitions, can have a non negligible impact on σ_s. Both the high values of σ_s (compared to the drift expectation)

and its pattern of variation (larger in assays with large fitness differences) support the conclusion that, even under very controlled and standardized conditions, cryptic environmental variation has a detectable impact on fitness measures.

Fitness transitivity: Population genetic models of selection usually consider fitness effects to be transitive between competing genotypes. In this view, fitness can be associated with a given genotype rather than being defined locally relative to particular competitors. (There are, of course, particular frequency dependent selection schemes that can generate non-transitive fitness measures (e.g. SINERVO and LIVELY 1996).) Methodologically, transitivity is also an important assumption in inferring allelic from genotypic fitness effects, as when using a marker to infer the effect of a mutation. Our experimental design allows us to test for departures from transitivity because we measured relative fitness in four combinations of genotypes pairs (see methods). Specifically, to test for transitivity, we need to make three estimates for a mutation: (1) the allelic cost of the marker, (2) the allelic effect of the mutation and (3) the epistasis between both. With three competitions, we have three equations and three unknowns. Thus, adding one competition adds one equation and provides a means to estimate a departure from consistency (*i.e.*, transitivity) among competition types. We found that τ, a parameter measuring deviations from transitivity, was not significantly different from zero for competitions involving the T63 and T121 mutants (LRT; Table 4), meaning that fitness was transitive. By contrast, τ was significantly different from zero for T103, but this departure was quite small ($\tau = -0.00171 \pm 0.0003$) and, more importantly, very small compared to the fitness differences measured in those competition assays (Table 4).

Allelic fitness and epistasis: To test for epistasis between our markers and the focal mutations, we decomposed genotypic fitness into the allelic effects of the marker and the mutation, and their interaction (epistasis). The expression of CFP was more costly than YFP (a 0.4% difference in the wild type) and the allelic effect of mutations was -1.2%, -1.7% and -0.3% for T63, T103, T121, respectively (Table 4). However, we detected significant differences between fitness effects of the same mutations when measured in the CFP and YFP backgrounds, indicating the existence of epistasis between the marker and the three individual mutations. Even though the strength of epistatic interactions was quite small, it could represent an important part of the genotypic selection coefficients. For instance, for

T63, epistasis was larger than the cost of the marker and represented 43% of the allelic mutational effect. For the two other mutations, the quantitative importance of epistasis was much smaller. A caveat to our interpretation of epistasis is that it is possible that we inadvertently introduced secondary mutations into genotypes during some step required for strain construction (see methods). Since we observed fitness differences with the three mutants we tested, the hypothesis of secondary mutation introduction supposes a very high rate of such mutations during P1 transduction.

DISCUSSION

In a large population, even mutations with very small fitness effects can play a role in the process of adaptation. However, studying them empirically is a significant practical challenge. Measurement error and drift obviously limit the precision of fitness measures that can be obtained experimentally. Is it possible to measure selection up to a limit imposed by the noise produced by sampling error and drift? If not, how close to this limit can we go? We addressed these questions by performing competition experiments in large *E. coli* populations (to minimize drift) and by tracking frequency changes using flow-cytometry to count marked cells (to minimize sampling error).

Our experiments are based on short-term batch cultures (6.6 generations). This design has several convenient features. First, it is a relatively simple experimental set up that can be massively replicated. Second, it reduces, though does not eliminate, the complication of newly arising mutations. Third, it entirely accounts for the effect of the marker. Last, it avoids the complication of using time series data.

Cryptic variation in *s*: A surprising, and we think important, result was that, for some competitions types, selection was variable across replicates, probably because of cryptic environmental variation to which the competing genotypes had different sensitivity. Although not empirically excluded, the alternative hypothesis that variation in estimates of *s* was caused by beneficial mutations spreading in a large number of the batch cultures, seems unlikely for two reasons: (1) such mutations would have to confer a very large benefit (unlikely to appear in a strain that has evolved in the same environment for 10,000 generations) and (2) adaptive mutations would increase var(*s*) in all competition types, not only in competitions (c) and (d). Such variation arose despite considerable effort to perform

all competitions in precisely controlled conditions. In an absolute sense, this variation was not large (although much larger than our precision), but it supports the idea that the effect of mutations can be strongly context-dependent. For instance, it is possible that if our experiment was performed in a different lab, the $\overline{s}$ and σ_s might be slightly different (because of differences in the average environment or in the magnitude of microenvironmental fluctuations, respectively). *A fortiori*, we expect σ_s to be even larger in environmentally heterogeneous natural conditions. These observations raise the question of whether selection coefficients should be described by only their mean values $\overline{s}$, or more appropriately by distributions (with two parameters $\overline{s}$ and σ_s), and consequently if mutations are appropriately described as beneficial, neutral or deleterious, since their effects are context-dependent, even within controlled laboratory environments. So far, population genetic models do not typically consider that *s* values are distributed, such that one mutation can have very different fates depending on its σ_s. For instance, the probability of fixation of a mutation with $\overline{s} = 0$ and $\sigma_s > 0$ will not be driven by drift only, as described by the neutral theory, but will depend on the environmental pattern responsible for $\sigma_s > 0$ (see *e.g.* (EWENS 1979)). In any case, much more attention should be paid to variable selection coefficients and their evolutionary impact. The experimental design and statistical analysis we propose here, offers an efficient and new approach to do that.

Epistasis with the marker: In competition experiments with microbes, the neutrality of the marker is always verified, however, the potential epistatic interactions between the marker and mutations is not usually systematically investigated. Controlling for this issue requires switching the markers between backgrounds (in order to perform complementary competition assays (DYKHUIZEN and HARTL 1980)) and a high level of precision. We found epistatic interactions between the inserted mutations and the fluorescent marker in all cases, suggesting that epistatic effects, though perhaps very small, may be common. Such interactions complicate measures of *s* because they require separating the MFE from the marker cost and the epistatic interactions between them. We note that if we had found that epistatic effects were of a similar size to (or larger than) the allelic effects it would have raised the concern that the compared strains may not have been isogenic. This was not the case in our experiment, nevertheless, we cannot formally exclude the possibility that

transformation and P1 transduction manipulations did not introduce any secondary mutations.

Transitivity: The assumption of fitness transitivity is made in most population genetic models that do not specifically include social effects or frequency dependence. This assumption has been evaluated on several occasions and in different organisms (BELL 2008b; DE VISSER and LENSKI 2002; GOODMAN 1979; PAQUIN and ADAMS 1983; RICHMOND *et al.* 1975). The main conclusion is that fitness tends to be transitive unless special social interactions are present. Like for any 'null hypothesis', it is, however, important to realize that the statistical power of an experiment gives an inherent limit to the detectable departure of transitivity. In our experiment, we tested this hypothesis and did not find consequential departures from transitivity in the genotypic fitness measures (Table 4). Given our high statistical power, this finding represents a strong internal check that our experimental results are robust. The fact that fitness effects are transitive is also an important result to simplify experiments: without the need to check for transitivity, only three types of competition need to be performed to estimate allelic effects and epistasis (instead of four in our design).

Precision of fitness measures and the statistics of selection: Although sampling error and drift can make it difficult to measure small fitness effects, replicated measures will tend not to significantly differ from each other. Consequently, a single fitness value can be legitimately attributed to a given genotype in a given environment and the precision of this estimate can be determined as the standard error of the mean fitness effect across replicates. New high-throughput counting methods alleviate limits due to sampling error and drift. Here, we show that such methods can reveal that replicated measures differ from one another for a given genotype in a given environment–i.e. that var(s) is significantly greater than the value expected by drift and sampling error alone. This situation challenges the simple concept of precision mentioned above. When confronted with this problem, one approach is to do "as usual" and neglect the observation that replicates differ. In this case, only providing a mean fitness effect and its standard error do not reflect the actual precision of the experiment. In particular, it fails to acknowledge that replicates may differ beyond this standard error. The second approach is to admit that replicates actually differ and represent different draws in a distribution of fitness effects, even if the environment is supposed constant. If selection coefficients are distributed, it is thus necessary to measure the mean

effect, but also the variance, and possibly higher moments (skewness, kurtosis, etc.) of the distribution of *s* values. The concept of precision in this case must incorporate estimates of these moments and their standard errors. This is the approach we have taken, introducing the use of a mixed model that allowed us to decompose sources of variation in frequency change (sampling error, drift, environmental variation of *s*).

We measure fitness based on frequency change as in classical population genetics (ARNASON and BARKER 1999; DYKHUIZEN and HARTL 1980), which differs from common practices in experimental evolution (CHEVIN 2010), where fitness is measured as a ratio of growth rates (e.g. LENSKI *et al.* 1991). We also fully use the information on the individual precision of fitness estimates (determined by the sampling effort: the number of colonies counted when plating, or the number of cells counted with flow cytometry), which is not usually reported with fitness competition experiments. Thus, we can discriminate between sampling error and other sources of variance across replicates (due to drift, variance in *s* etc.), which greatly enhances the information that can be extracted from the data. By considering these factors, we were able to measure mean and variance in selection coefficients down to a precision of 0.02% (Tables 3 and 4). Regarding mean selection, this precision represents a ~10-fold improvement over typical studies using flow cytometry (ALI and YANG 2006; LEE *et al.* 2009; LUNZER *et al.* 2002) and is comparable to the precision reached in (ZHU *et al.* 2005). More importantly, as explained above, the massive replication we used also provides a precise measure of the variance of selection coefficient σ_s (±0.02%).

Neutralist *vs*. selectionist: The neutral theory of molecular evolution proposes that the fate of many mutations is governed by the effect of drift (KIMURA 1983). The development of precise fitness measures was used in the neutralist / selectionist controversy on proteins in order to determine if allozymes differed in terms of selection (DYKHUIZEN and HARTL 1980). Today, this debate has shifted towards smaller fitness effects at the molecular level (KREITMAN 1996; NEI 2005). The analysis of sequence polymorphism provides different indirect ways to confront neutralist vs. selectionist expectations. For most mutations, it is usually thought that there is no alternative to resolve this question. Measuring *s* with ever increasing precision may start to change this perspective and may help to answer some of the key questions fueling this debate. As we have already seen in our study, mutations formerly considered as neutral (albeit in a slightly different medium, DM25 in (ELENA *et al.*

1998)), actually confer small, but significant, fitness effects. Importantly, even in an apparently constant environment, their effect is best understood as being distributed, which complicates straightforward application of discrete classifications (deleterious / neutral / beneficial), and which would have to be accounted for in theoretical expectations e.g. for the analysis of sequence polymorphism.

Sampling error, *de novo* beneficial mutations, and drift introduce elements of chance into fitness competitions, which can limit our ability to measure very small fitness effects. The effect of these factors can, however, be reduced. Sampling error can be dramatically decreased using flow cytometry, and the problem of the occurrence of new beneficial mutations and of drift is reduced by using very short term batch cultures and high replication. However, a remaining issue is the variation in selection due to micro-environmental variation across replicated cultures. While this would not be surprising if replicate measurements had been obtained from different environments (e.g. REMOLD and LENSKI 2001), that was not the case in our experiment in which all competitions were carried out in an environment that was kept as consistent as possible. Increased sampling effort, larger population sizes or longer lasting experiments are not likely to resolve this issue. It is thus unclear how far precision in fitness measurement can be improved. It all relies on understanding the sources of variation, and controlling them, whenever possible. Would it be possible to measure the fitness effects of synonymous mutations or mutations occurring in non-coding sequences? We cannot answer these questions yet, however, our method is a step in that direction and it will certainly help to bridge the gap between studies measuring *s* experimentally and studies inferring *s* from genetic sequences (see EYRE-WALKER and KEIGHTLEY 2007 for review).

Acknowledgements : We would like to thank E. Flaven, M.-P. Dubois for lab management, C. Duperray (IRB – Montpellier), C. Mongellaz (IGM – Montpellier) and the Montpellier RIO Imaging platform, for training RG to flow cytometry, and their help in designing the cytometer protocols, N. Le Meur for her help with the flowCore R package, L.M. Chevin, P.A. Gros, G. Martin and F. Rousset for fruitful discussions and insightful comments. We also thank two anonymous reviewers for useful comments. This work was supported by the ERC Starting Grant 'Quantevol' to T.L. and NSF (IOS-1022373) to T.F.C.

Figures and Tables

Step	Experiment	Description	Achievement
1	PCR	Introduction of the P_{A1} promoter in front of the CFP and YFP genes	PCR products called P_{A1}-CFP and P_{A1}-YFP
2	Cloning	Cloning P_{A1}-CFP and P_{A1}-YFP in pKD4*	plasmids called pKD4-CFP and pKD4-YFP
3	PCR	pKD4-CFP and pKD4-YFP were used as templates. Primers with 50 bases sequences homologous to the *E. coli rhaA* gene at their 5' ends were used.	PCR products called *rhaA*-CFP-Kan and *rhaA*-YFP-Kan
4	Transformation	The plasmid pKD46* was electroporated into the REL4548 recipient cells	REL4548 carrying pKD46
5	Transformation and homologous recombination	PCR products *rhaA*-CFP-Kan and *rhaA*-YFP-Kan were electroporated in REL4548 carrying pKD46.	REL4548 CFP-Kan^R and REL4548 YFP-Kan^R
6	Transformation	The plasmid pCP20* was electroporated into REL4548 CFP- Kan^R and REL4548 YFP-Kan^R	REL4548 CFP-Kan^R pCP20 and REL4548 YFP- Kan^R pCP20
7	Heat Shock	KanR cassette were excised from the REL4548 CFP- Kan^R and REL4548 YFP- Kan^R genomes	REL4548 CFP and REL4548 YFP

Table 1. Strain construction description. *from (DATSENKO and WANNER 2000)

Genotypes	Wild type CFP (wc)	Wild type YFP (wy)	Mutant CFP (mc)
Wild type YFP (wy)	$s_a = W_{wc}/W_{wy} - 1$		
Mutant CFP (mc)		$s_d = W_{mc}/W_{wy} - 1 + \tau$	
Mutant YFP (my)	$s_c = W_{my}/W_{wc} - 1$		$s_b = W_{mc}/W_{my} - 1$

Table 2. Selection coefficient expected in the different combinations of competition assays: (a) *wc*/*wy*, (b) *mc*/*my*, (c) *my/wc*, and (d) *mc*/*wy*. Each combination was performed for each mutant strain.

$(\times 10^{-3})$	$\bar{s}$	inf	sup	SE_{eq}	σ_s	inf	sup	SE_{eq}
Wild type (a)	-4.13	-3.74	-4.53	0.20	0.80	0.25	1.27	0.26
T63 (b)	1.16	1.58	0.74	0.21	0.59	0.00	1.18	0.29
T63 CFP (c)	-12.23	-11.93	-12.63	0.18	3.51	3.20	3.82	0.16
T63 YFP (d)	-8.28	-7.28	-6.28	0.29	1.49	1.02	1.94	0.23
T103 (b)	-6.48	-6.14	-6.81	0.17	0.00	0.00	0.60	0.15
T103 CFP (c)	-24.26	-24.10	-24.82	0.19	2.84	2.52	3.13	0.15
T103 YFP (d)	-15.55	-14.55	-13.55	0.24	1.11	0.71	1.60	0.22
T121 (b)	-3.60	-3.39	-3.80	0.10	0.00	0.00	0.44	0.11
T121 CFP (c)	-6.40	-5.92	-6.87	0.24	1.15	0.75	1.65	0.23
T121 YFP (d)	0.88	1.88	2.88	0.21	0.87	0.35	1.33	0.25

Table 3. Estimation of the mean ($\bar{s}$) and standard deviation (σ_s) of genotypic selection coefficients per generation in the different competition assays (code in first column). inf and sup indicates the inferior and superior support limits of the estimates. SE_{eq} gives a measure analogous to standard error and equals (sup-inf)/4. All figures are multiplied by 1000.

parameter	estimate $(\times\mathbf{10^{-3}})$	SE_{eq} $(\times\mathbf{10^{-3}})$	sign.
α	-4.13	0.14	***
β_{63}	-12.27	0.22	***
γ_{63}	5.28	0.22	***
τ_{63}	-0.13	0.30	ns
β_{103}	-17.89	0.22	***
γ_{103}	-2.23	0.22	***
τ_{103}	-1.71	0.30	**
β_{121}	-2.81	0.22	***
γ_{121}	0.55	0.22	*
τ_{121}	-0.48	0.30	ns

Table 4. Estimation of allelic selection coefficients per generation in the different competition assays (the subscript refers to the mutant T63, T103 or T121). α is the cost of the marker (bacteria expressing the fluorescent proteins CFP having a 0.4 % cost relative to those expressing YFP). β are the allelic effects of the three random mutations. γ are the epistasis between the random mutations and the marker. τ measures departure from transitive genotypic fitness. SE_{eq} gives a measure analogous to standard error and is a quarter of the support range. Significance (sign.) indicates whether the estimates are different from zero (LRT, *** < P-value 0.001, ** < P-value 0.01, ns: non significant). All figures are multiplied by 1000.

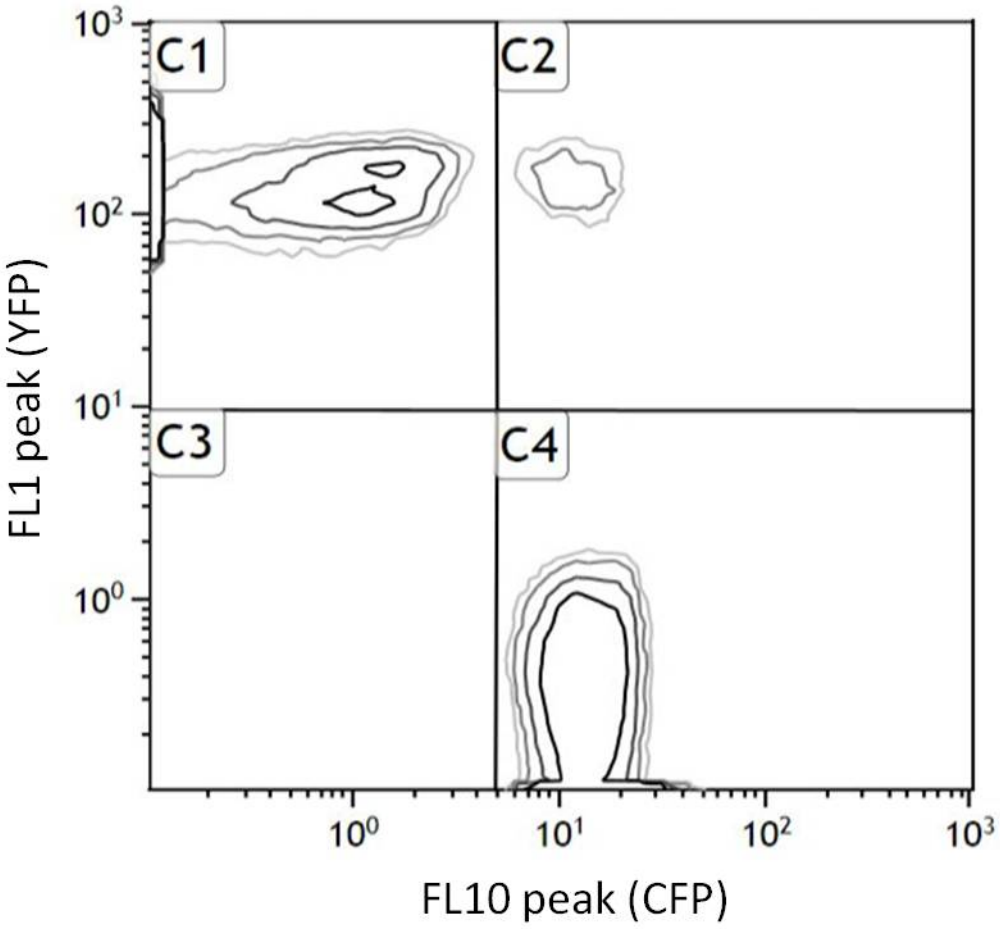

Figure 1. Measuring genotype frequencies with flow cytometry. The fluorescence of each bacterium was measured in the FL1 (YFP) and FL10 (CFP) channels. Here we show a representative contour plot. Quadrants represent thresholds delimiting the YFP (C1), CFP (C4), doubled marked (C2), and unmarked (C3) populations. In this example, populations are composed of YFP: 110718 (51.03%), CFP: 103996 (47.93%), doubled-marked: 2110 (0.97%), Unmarked 151 (0.07%) individuals.

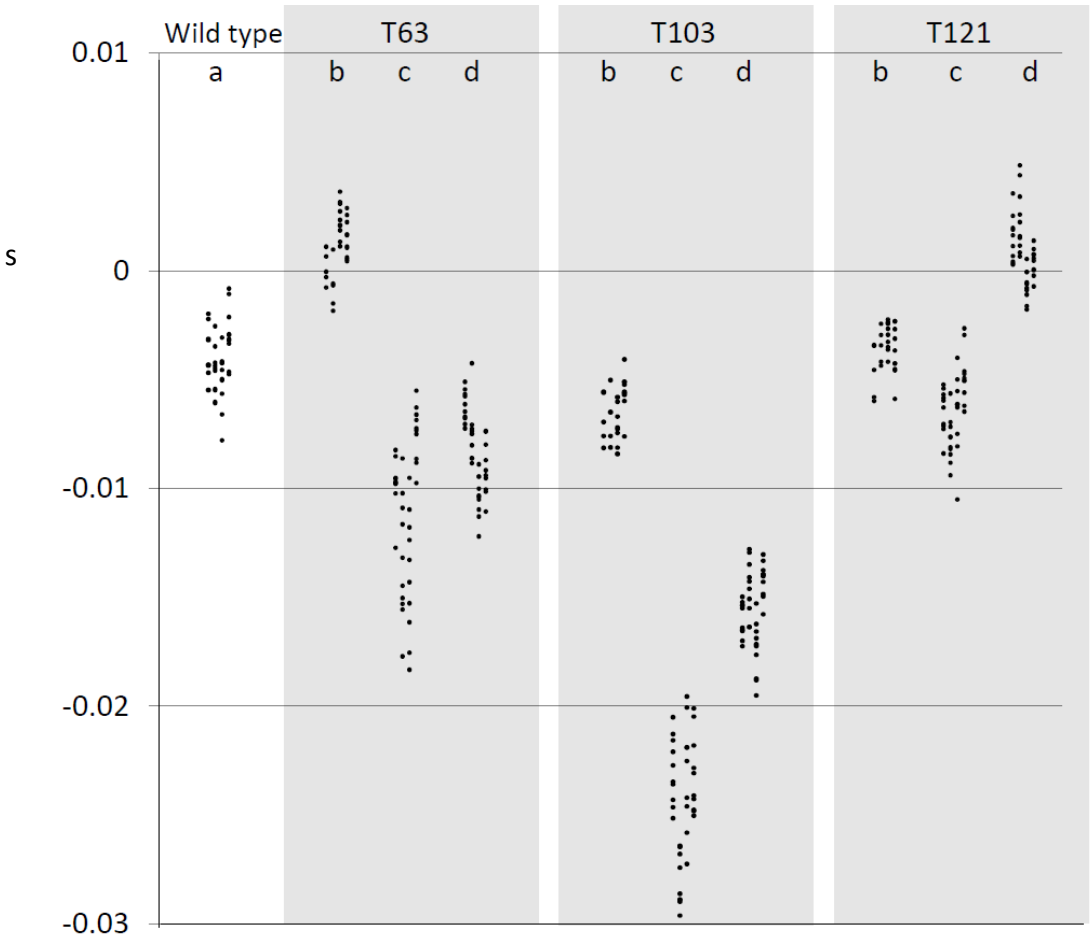

Figure 2. Selection coefficients (*s*) of the wild type (REL4548) and mutant strains (REL4548 T63, T103, or T121) measured in the (a) *wc*/*wy*, (b) *mc*/*my*, (c) *my*/*wc*, and (d) *mc*/*wy* competitions. Each point represents an *s* measure estimated from a single competition experiment. *s* measures are grouped in lines to show the variance between experiments performed at different dates.

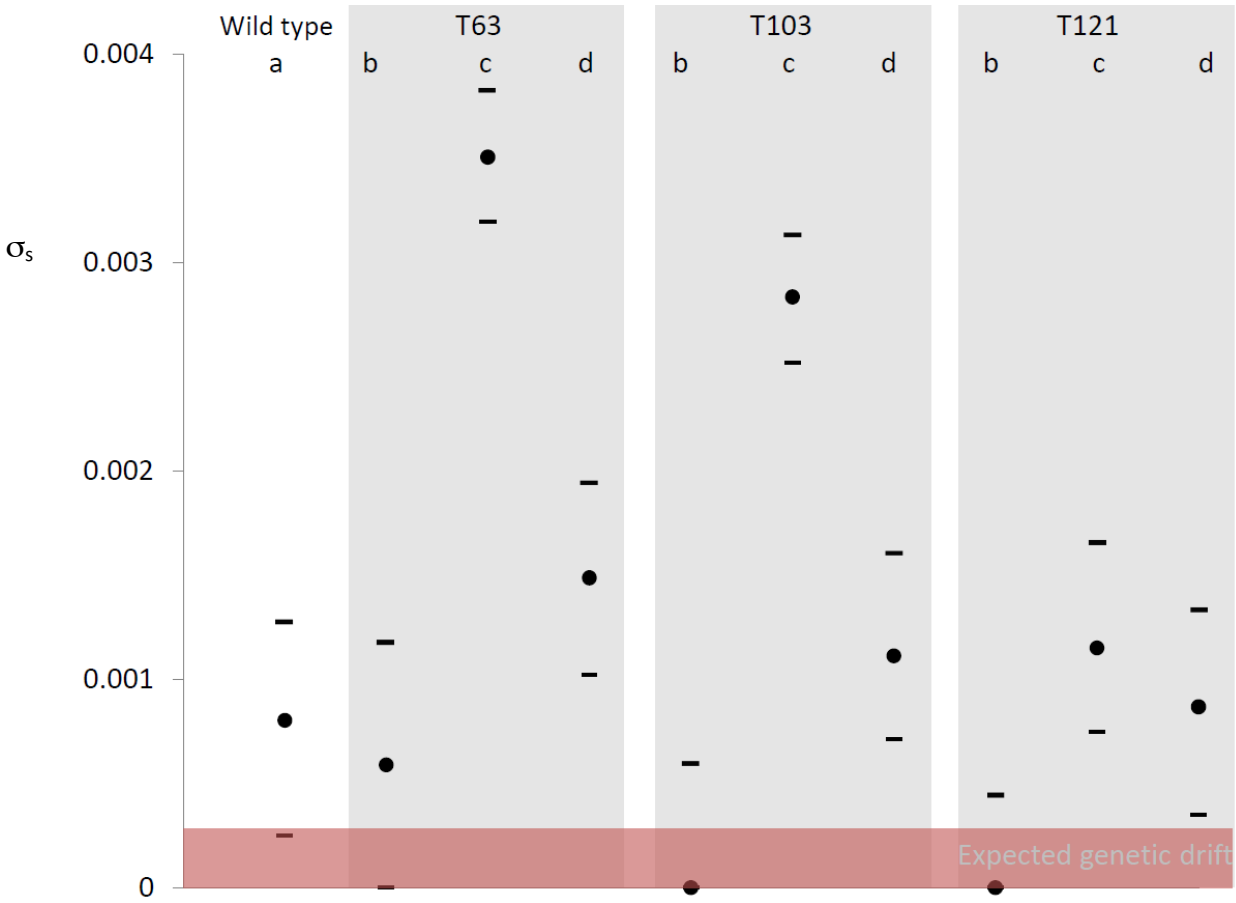

Figure 3. Variance in selection coefficients (σ_s estimate bold point ± support limits indicated by bars) of the wild type (REL4548) and mutant strains (REL4548 T63, T103, or T121) measured in the (a) *wc*/*wy*, (b) *mc*/*my*, (c) *my*/*wc*, and (d) *mc*/*wy* competitions. The predicted variation expected by genetic drift alone is represented by the shaded line at the bottom of the figure.

REFERENCES

ALI, A., and O. O. YANG, 2006 A novel small reporter gene and HIV-1 fitness assay. Journal of Virological Methods **133:** 41-47.

ARNASON, E., and J. BARKER, 1999 Analysis of selection in laboratory and field populations, pp. 182-203 in *Evolutionary genetics: from molecules to morphology*, edited by R. SINGH and C. KRIMBAS. Cambridge University Press, Cambridge.

ARNASON, E., and R. C. LEWONTIN, 1991 Perturbation-reperturbation test of selection vs. hitchiking ofthe 2 malor alleles of esterase-5 in *Drosophila pseudoobscura*. Genetics **129:** 145-168.

BALABAN, N. Q., J. MERRIN, R. CHAIT, L. KOWALIK and S. LEIBLER, 2004 Bacterial persistence as a phenotypic switch. Science **305:** 1622-1625.

BARRICK, J. E., D. S. YU, S. H. YOON, H. JEONG, T. K. OH *et al.*, 2009 Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature **461:** 1243-U1274.

BATAILLON, T., and M. KIRKPATRICK, 2000 Inbreeding depression due to mildly deleterious mutations in finite populations: size does matter. Genetical Research **75:** 75-81.

BEATTY, J., 1984 Chance and natural selection. Philosophy of Science **51:** 183-211.

BELL, G., 2008a *Selection*. Oxford Univ. Press, Oxford.

BELL, G., 2008b *Selection - The mechanism of Evolution*, NYC, USA.

BELL, G., 2010 Fluctuating selection: the perpetual renewal of adaptation in variable environments. Philosophical Transactions of the Royal Society B-Biological Sciences **365:** 87-97.

BOLLBACK, J. P., T. L. YORK and R. NIELSEN, 2008 Estimation of 2N(e)s from temporal allele frequency data. Genetics **179:** 497-502.

CARRASCO, P., F. DE LA IGLESIA and S. F. ELENA, 2007 Distribution of fitness and virulence effects caused by single-nucleotide substitutions in tobacco etch virus. Journal of Virology **81:** 12979-12984.

CHARLESWORTH, B., and D. CHARLESWORTH, 1998 Some evolutionary consequences of deleterious mutations. Genetica **103:** 3-19.

CHEVIN, L.-M., 2010 On measuring selection in experimental evolution. Biology Letters **Online publication doi: 10.1098/rsbl.2010.0580**.

CLARK, A., 1979 Effects of interspecific competition on the dynamics of a polymorphism in an experimental population of *Drosophila melanogaster*. Genetics **92:** 1315-1328.

DATSENKO, K. A., and B. L. WANNER, 2000 One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proceedings of the National Academy of Sciences of the United States of America **97:** 6640-6645.

DE VISSER, J. A. G. M., and R. LENSKI, 2002 Long-term experimental evolution in *Escherichia coli.* XI. Rejection of non-transitive interactions as cause of declining rate of adaptation. BMC Evolutionary Biology **2:** 19.

DOBZHANSKY, T., and O. PAVLOVSKY, 1957 An experimental study of interaction between genetic drift and natural selection. Evolution **11:** 311-319.

DOMINGO-CALAP, P., J. M. CUEVAS and R. SANJUAN, 2009 The fitness effects of random mutations in single-stranded DNA and RNA bacteriophages. PLoS Genet **5:** e1000742.

DYKHUIZEN, D., and D. HARTL, 1980 Selective neutrality of 6PGD allozymes in *E. coli* and the effects of genetic background. Genetics **96:** 801-817.

DYKHUIZEN, D. E., and D. L. HARTL, 1983 Selection in chemostats. Microbiological Reviews **47:** 150-168.

ELENA, S. F., L. EKUNWE, N. HAJELA, S. A. ODEN and R. E. LENSKI, 1998 Distribution of fitness effects caused by random insertion mutations in *Escherichia coli*. Genetica **102-3:** 349-358.
EWENS, W. J., 1979 *Mathematical population genetics*, Berlin, Heidelberg, New York.
EYRE-WALKER, A., and P. D. KEIGHTLEY, 2007 The distribution of fitness effects of new mutations. Nature Reviews Genetics **8:** 610-618.
FELSENSTEIN, J., 1976 The theoretical population genetics of variable selection and migration. Annual Review of Genetics **10:** 253-280.
FISHER, R. A., 1930 *The genetical theory of natural selection*. Oxford University Press, Oxford.
GOODMAN, D., 1979 Competitive hierarchies in laboratory drosophila. Evolution **33:** 207-219.
KIMURA, M., 1983 *The neutral theory of molecular evolution*. Cambridge university press, Cambridge, UK.
KLECKNER, N., J. BENDER and S. GOTTESMAN, 1991 Uses of transposons with emphasis on Tn10. Methods in Enzymology **204:** 139-180.
KREITMAN, M., 1996 The neutral theory is dead. Long live the neutral theory. BioEssays **18:** 678-683.
L'HERITIER, P., and G. TEISSIER, 1937a Elimination of the mutant forms in drosophila populations. Case of drosophila with the "ebony gene". Comptes Rendus Des Seances De La Societe De Biologie Et De Ses Filiales **124:** 882-884.
L'HERITIER, P., and G. TEISSIER, 1937b Elimination of the mutant forms in drosophila populations. The case of drosophila with the "bar gene". Comptes Rendus Des Seances De La Societe De Biologie Et De Ses Filiales **124:** 880-882.
LABBE, P., N. SIDOS, M. RAYMOND and T. LENORMAND, 2009 Resistance Gene Replacement in the Mosquito Culex pipiens: Fitness Estimation From Long-Term Cline Series. Genetics **182:** 303-312.
LEE, M. C., H. H. CHOU and C. J. MARX, 2009 Asymmetric, bimodal trade-offs during adaptation of methylobacterium to distinct growth substrates. Evolution **63:** 2816-2830.
LENORMAND, T., and M. RAYMOND, 2000 Analysis of clines with variable selection and variable migration. The American Naturalist **155:** 70-82.
LENORMAND, T., D. ROZE and F. ROUSSET, 2009 Stochasticity in evolution. Trends in Ecology & Evolution **24:** 157-165.
LENSKI, R. E., M. R. ROSE, S. C. SIMPSON and S. C. TADLER, 1991 Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. American Naturalist **138:** 1315 - 1341.
LEVIN, B. R., and D. E. ROZEN, 2006 Opinion - Non-inherited antibiotic resistance. Nature Reviews Microbiology **4:** 556-562.
LUNZER, M., A. NATARAJAN, D. E. DYKHUIZEN and A. M. DEAN, 2002 Enzyme kinetics, substitutable resources and competition: From biochemistry to frequency-dependent selection in *lac*. Genetics **162:** 485-499.
LYNCH, M., 1987 The consequences of fluctuating selection for isozyme polymorphisms in *Daphnia*. Genetics **115:** 657-669.
MANLY, B. F. J., 1985 *The statistics of natural selection*. Chapman and Hall Ltd, London.
MILLSTEIN, R. L., 2008 Distinguishing drift and selection empirically: "The great snail debate" of the 1950s. Journal of the History of Biology **41:** 339-367.
NEI, M., 2005 Selectionism and neutralism in molecular evolution. Molecular Biology and Evolution **22:** 2318-2342.

O'HARA, R. B., 2005 Comparing the effects of genetic drift and fluctuating selection on genotype frequency changes in the scarlet tiger moth. Proceedings of the Royal Society B-Biological Sciences **272:** 211-217.

OAKESHOTT, J. G., S. R. WILSON and J. B. GIBSON, 1983 An attempt to measure selection coefficients affecting the alcohol-dehydrogenase polymorphism in *Drosophila melanogaster* populations maintained on ethanol media. Genetica **61:** 151-159.

ORR, H. A., 2005 The genetic theory of adaptation: A brief history. Nature Reviews Genetics **6:** 119-127.

PAQUIN, C., and J. ADAMS, 1983 Frequency of fixation of adaptive mutations is higher in evolving diploid than haploid yeast populations. Nature **302:** 495-500.

PERIS, J. B., P. DAVIS, J. M. CUEVAS, M. R. NEBOT and R. SANJUAN, 2010 Distribution of fitness effects caused by single-nucleotide substitutions in bacteriophage ϕ1. Genetics **185:** 603-609.

REMOLD, S. K., and R. E. LENSKI, 2001 Contribution of individual random mutations to genotype-by-environment interactions in *Escherichia coli*. Proceedings of the National Academy of Sciences of the United States of America **98:** 11388-11393.

RICHMOND, R. C., M. E. GILPIN, S. PEREZSALAS and F. J. AYALA, 1975 Search for emergent competitive phenomena - Dynamics of multispecies drosophila systems. Ecology **56:** 709-714.

ROSENFELD, N., J. W. YOUNG, U. ALON, P. S. SWAIN and M. B. ELOWITZ, 2005 Gene regulation at the single-cell level. Science **307:** 1962-1965.

ROUSSET, F., 2004 *Genetic structure and selection in subdivided populations*. Princeton University Press, Princeton.

SACCHERI, I. J., F. ROUSSET, P. C. WATTS, P. M. BRAKEFIELD and L. M. COOK, 2008 Selection and gene flow on a diminishing cline of melanic peppered moths. Proceedings of the National Academy of Sciences of the United States of America **105:** 16212-16217.

SANJUAN, R., A. MOYA and S. F. ELENA, 2004 The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proceedings of the National Academy of Sciences of the United States of America **101:** 8396-8401.

SINERVO, B., and C. M. LIVELY, 1996 The rock-paper-scissors game and the evolution of alternative male strategies. Nature **380:** 240 - 243.

THATCHER, J. W., J. M. SHAW and W. J. DICKINSON, 1998 Marginal fitness contributions of nonessential genes in yeast. Proceedings of the National Academy of Sciences of the United States of America **95:** 253-257.

WILSON, S. R., J. G. OAKESHOTT, J. B. GIBSON and P. R. ANDERSON, 1982 Measuring selection coefficients affecting the alcohol-dehydrogenase polymorphism in *Drosophila melanogaster*. Genetics **100:** 113-126.

WOLFRAM RESEARCH, I., 2008 Mathematica, pp. in *Version 7.0*, Champaign, IL

ZHU, G. P., G. B. GOLDING and A. M. DEAN, 2005 The selective cause of an ancient adaptation. Science **307:** 1279-1282.

SUPPLEMENTARY MATERIALS

Insertions of the chromosomal fluorescent markers

The YFP and CFP genes were inserted at the *rhaA* locus of REL4548 using the one-step inactivation of chromosomal genes technique developed by Datsenko & Wanner (DATSENKO and WANNER 2000). pZA32-YFP and pZE1R-CFP - plasmids kindly provided by Dr Michael Elowitz - were used as templates.

The first step of this construction was to introduce the P_{A1} (bacteriophage λ promoter) upstream of the CFP and YFP genes. Because the YFP and CFP sequences (wild-type codons, developed by University of Washington Yeast Resource Center) only differ by 20 nucleotides (out of 717 bp), the same primers could be used for both PCRs. The primer P-BspHI-promA1-YC/FP for (109 bases) is composed of a 5' TCTC tail (increasing digestion efficiency by BspHI), a BspHI restriction site, the P_{A1} promoter containing a ribosome binding site (rbs) and a 20 bp sequence homologous to the YFP and CFP genes 5' extremity (see supplementary materials for primer sequences and PCR mix). The P-ClaI-YC/FP rev primer is composed of a 5' TCTC tail, a ClaI restriction site and a 20 bp sequence homologous to the YFP and CFP genes 3' extremity. The resulting PCR products were named P_{A1}-CFP and P_{A1}-YFP.

The second step consisted in cloning the P_{A1}-CFP and P_{A1}-YFP in pKD4 (DATSENKO and WANNER 2000). P_{A1}-CFP and P_{A1}-YFP were digested by BspHI and ClaI, and then gel purified. pKD4 was first digested by BspHI and ClaI, and then treated with shrimp alkaline phosphatase (SAP). The SAP catalyzes the release of 5'- and 3'-phosphate groups from DNA, precluding the ligation of the linearized plasmid with the remaining pKD4 BspHI-ClaI fragments. Finally, P_{A1}-CFP and P_{A1}-YFP were cloned in pKD4 by using a T4 ligase. Ligation products called pKD4-CFP and pKD4-YFP were electroporated into DH5-α λ-pir, and plated on LB plates supplemented with 50 μg/mL Kanamycin (LBA-Kan). Fluorescent colonies were purified by streaking on fresh LBA-Kan plates, and the next day, a colony was inoculated in liquid LB-Kan.

After isolation with Qiagen miniprep kit, pKD4-YFP and pKD4-CFP plasmids were used as templates in a second PCR in which the promoter sequence, the fluorescent genes and the Kanamycin cassette carried by pKD4, were amplified (see DATSENKO and WANNER 2000 for more details). The primers used in this second PCR - p-pKD4-RhaAH1-2863-2888 for (84

bases) and p-pKD4p2H2-1496-1477rev (79 bases) – contained either a BamHI or a SalI restriction site at their 5' extremities (restriction sites not used in this study), fifty bases homologous of the *E. coli rhaA* locus and the 21 or 27 bases homologous to the template at their 3' extremities. The resulting PCR products were electroporated into REL4548 competent cells carrying the thermosensitive plasmid pKD46 (DATSENKO and WANNER 2000) and transformed cells spread on LBA-Kan (incubation at 37°C in order to get rid of pKD46). Fluorescent recombinants were streaked on LBA-Kan for purification, and on LBA-Amp (50 μg/ml Ampicilin) to check for the loss of pKD46. PCRs were carried out on recombinant clones with two sets of primers (with a forward primer p-pKD4-2863-2888 specific to the insert, or p-RhaA-for specific to the *rhaA* gene, and the reverse primer RhaA-1262-1243-rev specific to the *rhaA* gene) to check that the insertion of CFP-Kan and YFP-Kan cassettes occurred at the right locus.

The excision of the kanamycin resistance cassette, corresponding to the last step of this cloning, was performed by the electroporation and induction of the plasmid pCP20 (DATSENKO and WANNER 2000). Transformants were first cultivated at 30°C on LBA-Amp plates. The flipase production (allowing the recombination of the FRT sequences surrounding the Kanamycin cassette) was induced by streaking colonies from the LBA-Amp plates, on LB plates at 42°C. To check for Kanamycin resistance cassette removal and loss of pCP20, colonies were streaked on LBA, LBA-Kan and LBA-Amp plates. PCRs on the *rhaA* locus were performed on 4548 CFP Kan^S Amp^S and 4548 YFP Kan^S Amp^S clones as well as on 4548, 4548 CFP-Kan and 4548 YFP-Kan. Both the Kan^S associated with a fluorescent phenotype and the size of the PCR products confirmed the removal of the kanamycin cassette and the insertion of the fluorescent genes.

Primer sequences

Name	Sequence	Construction step	Reference
P-BspHI-pA1-	TCTC**TCATGA**TTATCAAAAAGAGTA**TTGACT**TAAAGTCT AACCTATAG**GATACT**TACAGCCATCGAGAG*attaaagag*	PCR 1 Introduction of	This study

YC/FP for	*gagaaa*GGCGAAATGCGTAAAGGAGAAGAAC (109 bases) In blue : BspHI restriction site In black: P_{A1} In pink: Ribosome binding site In green, CFP/YFP homologous sequence	promoter pA1 upstream the CFP or YFP cassette	
P-ClaI-YC/FP rev	TCTCATCGATTTATTTGTATAGTTCATCCATGCC In blue : ClaI restriction site In green, CFP/YFP homologous sequence	PCR 1 Introduction of promoter pA1 upstream the CFP or YFP cassette	This study
p-pKD4-*rhaA*H1-2863-2888 for	GAGTCGACGGCGCTGCGCCAACTTGATCGTTTACCCGTTTCAATGCACTGCTGGCAGGGAAGCATTTATCAGGGTTATTGTCTC (84 bases) In blue : SalI restriction site In red: *rhaA* homologous sequence In green: pKD4-CFP (or pKD4-YFP) homologous sequence	PCR 2 Creating a linear fragment for homologous recombination	This study
p-pKD4p2H2-1496-1477rev	ACGGATCCACCACGCGGCAATGCGGTTGATAGAGGCATCGAAGAAGTCAAGGCCGATATCCTCCTTAGTTCCTATTCCG (79 bases) In blue : BamHI restriction site In red: *rhaA* homologous sequence In green: pKD4-CFP (or pKD4-YFP) homologous	PCR 2 Creating a linear fragment for homologou	This study

	sequence	s recombinati on	
p-*rhaA*-for	GACCACTC AACTGGAACAGGCC	PCR 3 Checking for CFP/YFP insertion in the *rhaA* locus	This study
p-pKD4-2863-2888 for	GAAGCATTTATCAGGGTTATTGTCTC	PCR 3 Checking for CFP/YFP insertion in the *rhaA* locus	This study
rhaA-1262-1243-rev	CCACGCTGGCTCAAAATCGC	PCR 3 Checking for CFP/YFP insertion in the *rhaA* locus	This study

1

2

PCR reactions:

PCR 1 - Introduction of promoter PA1 upstream the CFP or YFP cassette

Plasmid pZE1R-CFP or pZA32-YFP	1 μL
Primer P-BspHI-pA1-YC/FP for	2 μL
Primer P-ClaI-YC/FP rev	2 μL
Pfu 10X buffer	2 μL
dNTPs	1 μL
Pfu	0.5 μL
H_2O	11.5 μL

Program : Hybridization at 52°C for 1 min, elongation at 72°C for 2min, 25 cycles.

PCR 2 - Creating a linear fragment for homologous recombination

Plasmid pKD4-CFP (or pKD4-YFP)	2 μL
Primer p-pKD4-RhaAH1-2863-2888 for	1 μL
Primer pKD4p2H2-1496-1477rev	1 μL
Phusion 2X mastermix	10 μL
H_2O	6 μL

Program : Hybridization at 60°C for 1 min, elongation at 72°C for 3min, 25 cycles.

PCR 3 - checking CFP and YFP cassette insertions and Kanamycin cassette removal

Colony inoculated in the PCR mix, with a toothpick.

Primer forward	1 μL

Primer reverse	1 μL
Taq 10X buffer	2 μL
$MgCl_2$	1.5 μL
dNTP	1 μL
Taq Goldstar Red	0.2 μL
H_2O	13.3 μL

Program : Hybridization at 60°C for 1 min, elongation at 72°C for 3min 15 sec, 30 cycles.

Effect of "doublets" on coefficient of selection estimates

CY particles (C2), CC and YY doublet particles may pose a problem for accurate analysis of marker proportions. We have found that "singlet" and "doublet" particles can be nicely separated on a FSC TOF (time of flight) - FSC Peak plot. The frequency of YY (C1 region), CY (C2 region), and CC (C4 region) can thus be estimated. If cells were associated randomly in a doublet, when C and Y are initially introduce at a ratio of 1:1, we would expect to have C1 = 0.25, C2 =0.5 and C3 =0.25 within doublets. Most of the time, what we observe is C1 = 0.33, C2 = 0.33 and C3 = 0.33. This means that cells are not associated randomly in a doublet, but that they have more chance to be with a cell of the same color. Our hypothesis for such observation is that when entering stationary phase, a fraction of the daughter cell population might stay attached to each other. In Rang et al 2003, they observed that "*Cultures of the bacteria most affected by GFP exhibited a proportion of elongated cells, which suggests that GFP production could interfere with cell division in these strains*", which is another plausible explanation for our observation.

Now, the question is to evaluate the bias introduced by ignoring doublets. We can introduce some notation to be specific. There are several unknowns: the fraction of doublets (either CC, YY or CY, only the latter being measured in the C2 region). Let's denote it ϕ. Then there is the frequency of C bacteria (*p*). Last, we can consider that doublets may not form by the random association of two bacteria (e.g. two C bacteria may be more likely to form a

doublets than a pair of C and Y bacteria, as explained above). We can introduce a departure from random association *F* (akin to a departure from Hardy Weinberg proportion). It is then straightforward to compute the frequency of each type of particle (C, Y, CC, YY, CY). We measure frequency by the ratio (C+CC)/(C+CC+Y+YY), which introduces a bias from the true *p* value equal to

$$-(1-F)(1-p)p(1-2p)\phi + \mathrm{O}(\phi)^2.$$

This bias can be expressed in terms of ε, the proportion of the C2 population (defined as CY/(C+CC+CY+YY+Y). It is simply

$$-\epsilon(p-1/2)$$

which is very small, especially when *p* is close to ½. In our case, average initial frequency is 0.493±0.015 and ϵ < 1%, which corresponds to a bias close to 10^{-4}. The bias made on the frequency change, Δp, during the competition is different and it is this bias that is most relevant to estimating the strength of selection. Following the same approach and assuming F and ϕ are approximately constant during the competition, it is

$$\frac{1}{2}\Delta p(1-F)\phi$$

In other words the bias on the frequency change is proportional to itself. Expressing this in terms of ϵ and the intensity of selection per generation (with $\Delta p = s\ g\ pq$), we obtain $gs\epsilon pq$. The bias on selection coefficient is therefore equal to $s\epsilon$. Thus, it is always very small compared to *s*, provided the fraction of C2 population remains <<1. Correcting for doublets would be probably necessary when ϵ is greater than a few % (which, in our experiments, it is not). Under those circumstances, it would be important to have a clear experimental understanding of the origin of doublets particles (and measures of F and ϕ as we did). Last, it is important to underline that these slight biases do not introduce errors, so that estimations of variances remain unaltered.

Impact of de novo beneficial mutations during the assay

Beneficial mutations should have an implausibly high selection coefficient to impact fitness measures made in our experiments. The computation is different if we consider between week variance or within week (between replicates) variance.

Computation between weeks

The pre-cultures start from a frozen glycerol stock with a sample size of N_{gly} around 10^5. Importantly, all replicates within a week use the same pre-culture and only experiments performed in different weeks use different pre-cultures (*i.e.* different subsamples of the same glycerol stock). Seventeen generations later (around 10 generations during pre-culture, and seven more during a second round of pre-culture removing any trace of glycerol), the competition starts for 6.6 more generations. During the competition, the presence of a beneficial mutation with selection coefficient s_e may distort the frequency change measured in our experiment. If competitions always start with this beneficial mutation, little variance among competitions will be generated. On the contrary, variance in s among weeks will occur if the mutation is only present in some weeks. Hence, the worst case situation occurs when there is around ½ chances to sample the beneficial mutation from the glycerol. Computing this probability from a Poisson distribution, we find that the worst case is a frequency in the glycerol of around

$$p_{gly}=\text{Log}[2]/\ N_{gly}.$$

Starting from this initial frequency, the beneficial mutation will end up at the end of the preculture after 17 generations at a frequency

$$p_o = p_{gly}\ \text{Exp}[s_e\ 17]$$

(from logistic growth). Let's assume that the beneficial mutation occurred in a YFP cell. In our experiments, p_{yfp} –the frequency of the YFP cells—is always close to ½. During the

competition itself, the beneficial mutation will cause an extra change Δp_e in frequency of the YFP background

$$\Delta p_e = s_e p_o (1\text{-}p_{yfp}).$$

The bias per generation in the estimated selection coefficient is then

$$s_{bias} = \Delta p_e \,/\, (p_{yfp}(1\text{-}p_{yfp})\ 6.6).$$

Solving for the selection coefficient s_e causing a bias in selection at least as large as our precision (*i.e.* of the order of $s_{bias} = 2.10^{-4}$), we obtain $s_e = 0.33$. We are unaware of adaptive changes as high as this in the ecological situation of the growth environments we used. In particular, none has been found in the Lenski experiment. Moreover, our strain has already evolved 10,000 generations, which is the period where the strongest adaptive changes have already been fixed (Barrick et al., 2009). Indeed, the fitness of the relevant population increased by less than 30% in 20,000 generations following the time point at which REL4548 was isolated (Barrick et al., 2009).

Computation within weeks

Each replicated competition within a week starts from the same preculture and a large inoculum (10^6 cells). The worst case is when the beneficial mutation occurs early in half of the competitions. Since we sample $N_{start} = 5.10^5$ cells of a given competitor, it means that

$$p_{start}=\text{Log}[2]/\ N_{start}$$

Using the exact same computation as above, except that the beneficial has only 6.6 generation to change in frequency, we obtain s_e = 476, which is clearly unrealistic. Importantly, we observe inflated var(*s*) within weeks of the same order of magnitude as between weeks, so that it is quite clear that adaptive changes are unable to explain our results.