Campos, Zhao and Charlesworth

SI Appendix

1. Approximate BGS equations for a single gene without introns with no gene conversion

Effects of nonsynonymous sites without gene conversion.

First, consider the case with the same heterozygous selection coefficient t for each NS site. Let U be the deleterious haploid NS mutation rate for the gene in question ($U = 0.667 n_{ex} l_{ex} u$ with the model of nonsynonymous mutations in codons 1 and 2 alone described in the second section of the Materials and Methods), where u is the mean mutation rate per bp for the gene). Let P and Q be the fractions of the gene to the left and right, respectively, of a focal neutral site; M is the map length of the gene in Morgans, such that $M = r_c n_{ex} l_{ex}$. In the absence of gene conversion, Eq. 9 of (1) for a continuum of selected sites implies that the BGS effect for a neutral site at position P with selection coefficient t is:

$$E(P) \approx \frac{U(t + 2PQ\tilde{M})}{(2t + \tilde{M})} \left\{ \frac{1}{(t + P\tilde{M})} + \frac{1}{(t + Q\tilde{M})} \right\} \quad \text{(S1a)}$$

where the tilde over M denotes $M(1 - t)$.

The mean value of $E(P)$ for a gene is obtained by integrating Eq. S1A with respect to P over the interval (0, 1), giving:

$$\bar{E} \approx \frac{U}{(2t + \tilde{M})} \int_0^1 (t + 2PQ\tilde{M}) \left\{ \frac{1}{(t + P\tilde{M})} + \frac{1}{(t + Q\tilde{M})} \right\} \mathrm{d}P \quad \text{(S1b)}$$

The first component of the right-hand side of this equation is equivalent to:

$$\frac{U}{(2t + \tilde{M})} \int_0^1 \frac{\mathrm{d}P}{(t + P\tilde{M})} \mathrm{d}P = \frac{2Ut}{(2t + \tilde{M})\tilde{M}} \left[\ln\left(\frac{t + \tilde{M})}{t} \right) \right] \quad \text{(S2a)}$$

The second component can be written as:

$$\frac{4U\tilde{M}}{(2t + \tilde{M})} \int_0^1 \frac{(P - P^2)\, \mathrm{d}P}{(t + P\tilde{M})} \quad \text{(S2b)}$$

We have:

$$\frac{4U\tilde{M}}{(2t+\tilde{M})}\int_0^1 \frac{P\,\mathrm{d}P}{(t+P\tilde{M})} = \frac{4U}{(2t+\tilde{M})}\left[1-\frac{t}{\tilde{M}}\ln\left(\frac{t+\tilde{M}}{t}\right)\right] \qquad \text{(S3)}$$

and

$$\frac{4U\tilde{M}}{(2t+\tilde{M})}\int_0^1 \frac{P^2\,\mathrm{d}P}{(t+P\tilde{M})} = \frac{4U}{(2t+\tilde{M})\tilde{M}}\left[\tfrac{1}{2}(\tilde{M}-2t)+\frac{t^2}{\tilde{M}}\ln\left(\frac{t+\tilde{M}}{t}\right)\right] \qquad \text{(S4)}$$

Adding the separate terms together yields:

$$\bar{E} \approx \frac{2U}{\tilde{M}}\left\{1-\frac{t}{\tilde{M}}\ln\left(\frac{t+\tilde{M}}{t}\right)\right\} \qquad \text{(S5)}$$

When $M(1-t) < t$, the logarithm can be expanded in powers of $M(1-t)/t$, so that the leading terms in mean E are given by:

$$\bar{E} \approx \frac{U}{t}\left(1-\frac{2\tilde{M}}{3t}\right) \qquad \text{(S6)}$$

When $M(1-t) < t$, this is a decreasing function of t, showing that BGS within a single gene can in principle generate a negative relation between the level of selective constraint on the gene and the level of synonymous site diversity.

Approximate net effect of BGS due to nonsynonymous sites for a randomly placed neutral site in a gene without introns. Let the gene be $n_{ex}l_{ex}$ ($> d_g$) basepairs in length. We consider the kth site counted from the right end of this sequence. The net effect of BGS contributed by the nonsynonymous sites to the right of this site can be written as:

$$E(k) = 0.667u\int_0^k \frac{t\,dx}{[t+r(x)(1-t)]^2} \qquad \text{(S7)}$$

where x represents distance in bp, approximated by a continuous line, and $r(x)$ is the net rate of recombination over distance x, given by Eq. 2 of the Materials and Methods. Since this is an exponential function of x, the integral cannot be evaluated analytically.

A simple model of the effect of gene conversion is to approximate $1 - \exp(-d_{ij}/d_g)$ by d_{ij}/d_g, which is accurate when $d_{ij} << d_g$. This implies that r_{ij} can be replaced by $(r_c + g_c)d_{ij}$, i.e. gene conversion behaves similarly to crossing over, so that M can replaced with $(r_c + g_c)n_{ex}l_{ex}$. Alternatively, when $d_{ij} >> d_g$, the exponent is negligible and the expected value of the rate of gene conversion from Eq. 2 is approximately $g = d_g g_c$. The product of this term and $1 - t_i$ can simply be added to t_i in the denominator of the basic equation for the effect on BGS of a single selected site i, yielding $E_i = u_i t_i / [t_i + (g + r_c d_{ij})(1 - t_i)]^2$.

A better alternative to both of these methods is to use the first approximation when $d_{ij} \le d_g$ and the second approximation when $d_{ij} > d_g$; this is the 'mixed model' of gene conversion:

$$\hat{r}(x) = \begin{cases} (r_c + g_c)x & \text{if } x \le d_g \\ r_c x + g_c d_g & \text{otherwise} \end{cases} \qquad \text{(S8)}$$

In both cases, the effect of gene conversion is overestimated, so that the effect of BGS will be underestimated when this 'mixed' approximation is used. Applying these relations to Eq. S7, we obtain:

$$E(k) \approx 0.667ut \int_0^k \frac{dx}{[t + r(x)(1-t)]^2} = \begin{cases} \dfrac{0.667ut}{(1-t)(r_c + g_c)}\left[\dfrac{1}{t} - \dfrac{1}{t + (r_c + g_c)(1-t)k}\right] & k \le d_g \\ \dfrac{0.667u}{(1-t)(r_c + g_c)} - \dfrac{0.667ut}{(1-t)r_c}\dfrac{1}{t + [g_c d_g + r_c k](1-t)} & \\ \quad + \dfrac{0.667utg_c}{(1-t)r_c(r_c + g_c)}\dfrac{1}{t + d_g(r_c + g_c)(1-t)} & k > d_g \end{cases} \qquad \text{(S9)}$$

We then integrate $E(k)$ from $k = 0$ to $n_{ex}l_{ex}$, which gives its mean over all k values:

$$E(k) \approx 0.667ut\int_0^{n_{ex}l_{ex}} E(k)\,\mathrm{d}k$$
$$= \frac{0.667un_{ex}l_{ex}}{(1-t)(r_c+g_c)} + \frac{0.667g_c(n_{ex}l_{ex}-d_g)ut}{\left[(1-t)r_c(r_c+g_c)\right]\left[t+d_g(r_c+g_c)(1-t)\right]}$$
$$-\frac{0.667ut}{\left[(1-t)(r_c+g_c)\right]^2}\ln\left[\frac{t+d_g(r_c+g_c)(1-t)}{t}\right] - \frac{0.667ut}{\left[(1-t)r_c\right]^2}\ln\left[\frac{t+(r_cn_{ex}l_{ex}+g_cd_g)(1-t)}{t+(r_cd_g+g_cd_g)(1-t)}\right] \quad \text{(S10a)}$$

The same procedure can be applied to the effects of nonsynonymous sites to the right of the initial synonymous site, so that the final value is twice the above expression. To take the mean over all sites, we divide the final values by the total number of sites in the gene ($n_{ex}l_{ex}$). Noting that the deleterious mutation rate for the gene is $U = 0.667\, n_{ex}l_{ex}u$, after some manipulation this yields :

$$\bar{E} \approx \frac{2U}{\tilde{M}}\left\{\frac{1}{(1+r)} + \frac{rt(1-d_gr_c/M)}{(1+r)[t+d_g(r_c+g_g)(1-t)} - \frac{t}{\tilde{M}(1+r)^2}\ln\left[\frac{t+d_g(r_c+g_c)(1-t)}{t}\right]\right.$$
$$\left. -\frac{t}{\tilde{M}}\ln\left[\frac{t+(g_c+M)(1-t)}{t+d_g(r_c+g_c)(1-t)}\right]\right\} \quad \text{(S10b)}$$

where $r = g_c/r_c$. When $g_c = 0$, this expression reduces to Eq. S5.

Contribution of large effect mutations. Data from quantitative genetics analysis of mutational effects on fitness suggest that there is a contribution of deleterious mutations with larger effects than those estimated using DFE-alpha, representing mutational events that are distinct from those included in the rate U (2). It is a reasonable approximation to assume a fixed selection coefficient t_l for such mutations. If the proportion of all mutations that have large effects is p_l, the mutation rate to large mutations is $U_l = p_lU/(1 - p_l)$. We can then replace U and t in Eq. S10b with U_l and t_l, and add the resulting terms to Eq. S10b. This addition was used in all the models, unless otherwise stated.

Integral model of the effects of UTR sites in the absence of introns. We now examine the effects of untranslated regions (UTRs) of a gene, which also exert BGS effects on its synonymous sites. Assume that the length of the 5′ UTR is l_5, while the length of the 3′ UTR is l_3. Let u_u and t_u be the mutation and selection parameters for UTRs, corresponding to u and t for nonsynonymous sites in the previous section. We

assume that all sites in UTRs can exert BGS effects, which maximises the effect of BGS. Using the same procedure as above, the BGS effect contributed by the 3′ UTR for the *l*th site from the right-hand end of an exon is:

$$E_3(l) = \int_l^{l+l_3} \frac{u_u t_u}{\left[t_u + r(x)(1-t_u)\right]^2} \mathrm{d}x$$

$$= \begin{cases} \dfrac{u_u t_u}{(1-t_u)(r_c+g_c)}\left[\dfrac{1}{t_u+(r_c+g_c)(1-t_u)l} - \dfrac{1}{t_u+(r_c+g_c)(1-t_u)(l+l_3)}\right] & 0 \le l < \max\{0, d_g - l_3\} \\ \dfrac{u_u t_u}{(1-t_u)(r_c+g_c)}\dfrac{1}{t_u+(r_c+g_c)(1-t_u)l} - \dfrac{u_u t_u}{(1-t_u)r_c}\dfrac{1}{t_u+\left[g_c d_g + r_c(l+l_3)\right](1-t_u)} & \max\{0, d_g - l_3\} \le l < d_g \\ \dfrac{g_c u t_u}{(1-t_u)r_c(r_c+g_c)}\dfrac{1}{t_u+(r_c+g_c)(1-t_u)d_g} & \\ \dfrac{u_u t_u}{(1-t_u)r_c}\left\{\dfrac{1}{t_u+\left[g_c d_g + r_c l\right](1-t_u)} - \dfrac{1}{t_u+\left[g_c d_g + r_c(l+l_3)\right](1-t_u)}\right\} & l \ge d_g \end{cases}$$

(S11a)

where the mixed model of gene conversion was used.

Integrating this expression over all synonymous sites, the total BGS effect due to a 3′ UTR is:

$$E_3 = \int_0^{n_{ex}l_{ex}} E_3(l)\,\mathrm{d}l$$

$$= \begin{cases} \dfrac{u_u t_u}{(1-t_u)^2(r_c+g_c)^2}\ln\left[\dfrac{t_u+(r_c+g_c)d_g(1-t_u)}{t_u}\right] - \dfrac{u_u t_u}{(1-t_u)^2 r_c^2}\ln\left\{\dfrac{t_u+\left[g_c d_g + r_c(n_{ex}l_{ex}+l_3)\right](1-t_u)}{t_u+\left[g_c d_g + r_c l_3\right](1-t_u)}\right\} & \\ + \dfrac{u_u t_u}{(1-t_u)^2 r_c^2}\ln\left\{\dfrac{t_u+\left[g_c d_g + r_c n_{ex}l_{ex}\right](1-t_u)}{t_u+(r_c+g_c)d_g(1-t_u)}\right\} + \dfrac{g_c u_u t_u}{(1-t_u)r_c(r_c+g_c)}\dfrac{d_g}{t_u+(r_c+g_c)(1-t_u)d_g} & d_g < l_3 \\ \dfrac{u_u t_u}{(1-tu)^2(r_c+g_c)^2}\ln\left[\dfrac{t_u+(r_c+g_c)l_3(1-t_u)}{t_u}\right] - \dfrac{u_u t_u}{(1-t_u)^2 r_c^2}\ln\left[\dfrac{t_u+\left[g_c d_g + r_c(n_{ex}l_{ex}+l_3)\right](1-t_u)}{t_u+\left[g_c d_g + r_c d_g\right](1-t_u)}\right] & \\ + \dfrac{u_u t_u}{(1-t_u)^2 r_c^2}\ln\left\{\dfrac{t_u+\left[g_c d_g + r_c n_{ex}l_{ex}\right](1-t_u)}{t_u+(r_c+g_c)d_g(1-t_u)}\right\} + \dfrac{g_c u_u t_u}{(1-t_u)r_c(r_c+g_c)}\dfrac{l_3}{t_u+(r_c+g_c)(1-t_u)d_g} & d_g \ge l_3 \end{cases}$$

(S11b)

A similar expression applies to the BGS effect of a 5′ UTR, with 5 replacing 3.

If $d_g > l_3, l_5$, as is the case for *Drosophila*, the mean effect of BGS from the two UTRs on a synonymous site is given by:

$$\frac{E_3+E_5}{n_{ex}l_{ex}}=\frac{U_u t_u}{(1-t_u)^2(1+r)^2 M_{ex}(M_3+M_5)}\ln\left\{\frac{\left[t_u+M_3(1+r)(1-t_u)\right]\left[t_u+M_5(1+r)(1-t_u)\right]}{t_u^2}\right\}$$
$$-\frac{U_u t_u}{(1-t_u)^2 M_{ex}(M_3+M_5)}\ln\left\{\frac{\left[t_u+(g_c d_g+M_{ex}+M_3)(1-t_u)\right]\left[t_u+(g_c d_g+M_{ex}+M_5)(1-t_u)\right]}{\left[t_u+(g_c d_g+M_{ex})(1-t_u)\right]^2}\right\}$$
$$+\frac{g_c U_u t_u}{(1-t_u)r_c(1+r)M_{ex}\left[t_u+(r_c+g_c)(1-t_u)d_g\right]} \qquad \text{(S12)}$$

where the total mutation rate for UTRs is $U_u = (l_3 + l_5)u_u$, and the total map lengths for exons and UTRs are $M_{ex} = r_c n_{ex} l_{ex}$, $M_3 = r_c l_3$ and $M_5 = r_c l_5$.

2. Determining the distributions of selection coefficients

To predict BGS effects for an assigned pair of values of β and ω_{na}, as was done to produce the results shown in Figure 2, we employed a modification of the approach of (3) and (4). This uses the level of non-adaptive NS divergence to obtain an estimate of the mean of the DFE, given a value of β. For this purpose, we used the formula of (5) as modified by (6) in their Eq. 23. This expression relates divergence due to the fixation of slightly deleterious mutations (relative to neutral divergence) to the parameters of a gamma distribution of t. Using this result, the mean of $\gamma = 4N_e t$ is given by:

$$\ln(\gamma) \approx [(1+\beta)\ln(\beta) + \ln[\zeta(1+\beta) - \ln(\omega_{na})]/\beta \qquad \text{(S13a)}$$

where $\zeta(1+\beta)$ is Riemann's zeta function:

$$\zeta(1+\beta) = \sum_{i=1}^{\infty} i^{-(1+\beta)} \qquad \text{(S13b)}$$

This method was also employed when estimating BGS effects from the population genomic data, as described in the next section, because it was found to provide more stable distributions of the bootstrapped values of γ than the direct estimates of γ obtained from DFE-alpha, probably reflecting inaccuracies in numerical integration over the gamma distributions. In this case, estimates of the relevant parameters are needed for each bin of K_A values. As described in the Materials and Methods, the DFE-alpha program (7) provided estimates of the shape (β) and scale parameters of a gamma

distribution for each bin, as well as estimates of ω_a and ω_{na}. The value of γ for a given bin was then calculated from β and ω_{na}, using Eqs. S13.

For this purpose, K_S values were first corrected for the effect on divergence of selection on codon usage at synonymous sites, using the mean *Fop* value for each bin to estimate the R parameter in Eq. S22 in section 5 below, which predicts the ratio of K_S to the corresponding neutral value. We then divided each observed K_S value by R, which yields the value of the neutral divergence corresponding to K_S. Adjusted values of ω_{na} for NS sites were obtained by multiplying the unadjusted values by R. A similarly adjusted estimate of the mean of K_S over all bins was used to determine the divergence time in Eqs. S17, which were used to estimate the substitution rates of selectively favorable NS and UTR mutations.

3. Integration of the BGS equations over the distribution of selection coefficients

In order to obtain estimates of the effects of BGS when there is a distribution of selection, nuumerical integration of the BGS equations over an assumed distribution of selection coefficients, for given values of the mutation and recombination parameters, was carried out. For the results shown in Figure 2, we assumed a gamma distribution of the scaled selection coefficient, $\gamma = 4N_e t$, with a fixed scale parameter β; this has convenient properties and provides a good fit to the data from the Rwandan population of *D. melanogaster* (8). We used a set of fixed values of the non-adaptive divergence parameter ω_{na}, and estimated the corresponding mean values of γ from each pair of ω_{na} and β values by the method described in section 2 above
(Eqs. S13). For a given value for N_e (10^6 for the *D. melanogaster* population considered here (9)), the probability density of t for a given pair of values of β and ω_{na} was obtained from a gamma distribution.

We removed the portion of the distribution of t for which the scaled selection coefficient ($\gamma = 4N_e t$) was less than a threshold value, γ_c, of order 1, since the standard BGS formula is an overestimate when mutations are so weakly selected that drift becomes important (1). The selection coefficient corresponding to γ_c is denoted by t_c. By comparing the simulation results in Table 2 of (1) to their Eq. 9, we found that removal of the portion of the gamma distribution below $4N_e t$ =5 led to accurate predictions of the simulated effects of BGS in a finite population. We therefore used γ_c = 5 for the BGS predictions described in this paper. Use of the more stringent γ_c = 1 resulted in small quantitative differences to the results, with a slight strengthening of the net BGS effect.

This had little effect on the estimates of the parameters for positive selection, which were estimated as described below.

For the summation method, a value of t was drawn from the truncated distribution for each selected site i and applied to Eq. 1 of the Materials and Methods. The sum over i was then taken to get E_j, using an assumed value of the mutation rate per bp, u. This was done for all values of j, and the mean of E_j over all sites was then determined. This was repeated 500 times to obtain a mean value of E, $\bar{E}$, for a synonymous site located at a random position in the gene. The BGS predictions using the integral approximation were obtained from the 'mixed model' of gene conversion, described by Eqs. S10b and S12 above, integrating over the assumed truncated gamma distributions for NS and UTR sites, respectively.

4. The effects of selective sweeps on diversity

Consider first the spread of a favorable mutation at nonsynonymous site i, which arose in a haplotype carrying a particular allele at a neutral site j. Following (10), the net change in the frequency q_j of the neutral allele between the start and end of the sweep in a large population is given by:

$$\Delta q_j \approx (1-q_{j0}) q_{i0}^{2r_{ij}/s} \int_{q_{i0}}^{1} \left(\frac{[1-q_{it}]}{q_{it}} \right)^{2r_{ij}/s} \mathrm{d}q_{it} \qquad \text{(S14a)}$$

where $q_{j\,t}$ and $q_{i\,t}$ are the frequencies of the neutral and selectively favorable alleles, respectively in generation t after the initiation of the sweep; r_{ij} is the recombination frequency between the two sites, and s is the selection coefficient in favor of homozygotes for the beneficial allele (semidominance and weak selection are assumed). An elementary derivation of this and the following results is given in (11), pp.410-411.

Favorable mutations that arise on the alternative background with respect to the neutral locus contribute a change in allele frequency of the chosen allele at the neutral locus of:

$$\Delta \tilde{q}_j \approx -q_{j0} q_{i0}^{2r_{ij}/s} \int_{q_{i0}}^{1} \left(\frac{[1-q_{it}]}{q_{it}} \right)^{2r_{ij}/s} \mathrm{d}q_{it} \quad \text{(S14b)}$$

The probabilities of these two alternative events are $q_{j\,0}$ and $1-q_{j\,0}$, respectively.

The integral in Eqs. S14a and S14b is an incomplete beta function and cannot be

evaluated analytically, but it is approximately equal to one when $r_{ij} << s$, as is required for a significant effect of the sweep on site *j*. We show in section 6 below that this assumption is likely to yield accurate results for cases of biological interest. In order to correct for stochastic losses of mutations arising at initial frequencies of $1/(2N)$, $q_{i\,0}$ is equated to the ratio of $1/(2N)$ to the fixation probability of the favorable allele, $(N_e s/N)$, giving $q_{i\,0}=1/(2N_e s)$ (12). These two assumptions give the change in allele frequency from Eq. S14a as:

$$\Delta q_j \approx (1-q_{j0})\gamma_a^{-2r_{ij}/s} \qquad \text{(S14c)}$$

where $\gamma_a = 2N_e s$ is the scaled selection coefficient for the favorable allele (a factor of 2 not 4 is used, since *s* is the selection coefficient for homozygotes).

The expected reduction in diversity caused by a sweep, relative to the neutral values, can be obtained by averaging over the two possible allelic states at the neutral locus and using Eq. S14c. It is given by:

$$\frac{\Delta\pi}{\pi_0} \approx -\gamma_a^{-4r_{ij}/s} \qquad \text{(S14d)}$$

This equation can also be derived by arguments based on coalescent theory (13, 14). Note that an incorrect version was used by (15), which may affect their inferences concerning the strength of selection.

As noted in (16), if the reasonable assumption is made that the time over which the selectively favorable allele spreads to fixation is much shorter than the expected neutral pairwise coalescent time, $2N_e$, this relative change in diversity is equivalent to the probability that there was no recombination during the sweep, so that the neutral site experienced a reduction of coalescent time to zero, as opposed what happens when recombination events occurred during the sweep that allow neutral sites with a standard coalescent time to be introduced into haplotypes carrying the favorable allele.

Following (16) and subsequent authors, e.g. (14), if we consider recurrent sweeps occurring over all nonsynonymous sites in a gene, at a rate ν_a per nonsynonymous site per generation, and assume that ν_a is sufficiently small that each sweep exerts its effect independently of the others, the net probability of an effectively instantaneous coalescent event induced by selection at neutral site *j* is given by:

$$P_{ca\,j} = \nu_a \sum_i \gamma_a^{-4r_{ij}/s} \qquad \text{(S15a)}$$

A similar expression can be written for the effects of selective sweeps in the UTRs, where for simplicity we assume that the 3´and 5´ UTRs have the same selection coefficient, which may of course differ from that for nonsynonymous sites. We can simply replace ν_a with a corresponding rate of sweeps per each UTR sites (ν_u), and γ_a with a scaled selection coefficient γ_u, yielding a rate of selectively induced coalescent events of:

$$P_{cu\,j} = \nu_u \sum_{i'} \gamma_u^{-4r_{i'j}/s'} \qquad \text{(S15b)}$$

where the primes denote UTR sites and their parameters.

Provided that $P_{ca\,j}$ and $P_{cu\,j}$ are << 1, the selectively induced coalescent events and standard neutral coalescent events at site *j* can be regarded as competing exponential processes (16), with rates $P_{cs\,j} = P_{ca\,j} + P_{cu\,j}$ and $1/(2N_e)$, respectively. The net rate of coalescence is then given by the sum of these terms. Taking the reciprocal of this rate, the expected pairwise coalescent time at site, relative to the neutral value in the absence of sweeps, $2N_e$, is given by:

$$\frac{T_j}{2N_e} = \frac{1}{1 + 2N_e P_{sc\,j}} \qquad \text{(S16a)}$$

The effects of BGS can be included by assuming that it acts independently of sweeps, reducing the effective population size at site *j* from $2N_e$ to $2N_e B_j$, where B_j is given by Eq. 1 of the Materials and Methods or one of the approximations in section 1 above (14, 17, 18). Eq. S16a can then be replaced with:

$$\frac{T_j}{2N_e} = \frac{1}{B_j^{-1} + 2N_e P_{sc\,j}} \qquad \text{(S16b)}$$

Assuming the infinite sites model of neutral variability (19), the ratio of neutral diversity at a synonymous site *j* (π_j) to its expected value in the absence of selection (π_0) is thus given by:

$$\frac{\pi_j}{\pi_0} = \frac{1}{B_j^{-1} + 2N_e P_{sc\,j}} \qquad \text{(S16c)}$$

For a two-species comparison, we can write ν_a and ν_u in equations (S15a) and (S15b) as:

$$\nu_a = \alpha_a K_A / (2t_{div}) \qquad \text{(S17a)}$$

$$\nu_u = \alpha_u K_U / (2t_{div}) \qquad \text{(S17b)}$$

where t_{div} is the divergence time between the two species being compared, K_A and K_U are the nonsynonymous site and UTR site divergences, and α_a and α_u are the corresponding proportions of mutations fixed by positive selection. For substitutions along a single lineage, the factor of 2 is omitted.

Under neutrality, $2ut_{div} = K_S$ for a two species comparison, and $ut_{div} = K_S$ for a single lineage, so that t_{div} can be estimated from estimates of u and K_S (a correction for the effect on K_S of selection on codon usage is described in section 5 below). Under the infinite sites model, neutral nucleotide site diversity is proportional to the coalescent time, so that Eq. S16c can be rearranged to give the deviation between the predicted and observed values of the negative of the logarithm of the ratio of synonymous diversity to its value in the absence of selection. For the two-species comparion, we have:

$$dev_j = -\ln\left(\frac{\pi}{\pi_0}\right) + \ln\left(\frac{1}{B_j^{-1} + \alpha_a K_A (2t_c)^{-1} \sum_i \gamma_a^{-4r_{ij}/s} + \alpha_u K_U (2t_c)^{-1} \sum_{i'} \gamma_u^{-4r_{i'j}/s'}}\right) \qquad \text{(S18)}$$

where t_c is the divergence time in coalescent time units ($t_c = t_{div}/2N_e$) for the two-species comparison. For a single lineage, the factor of 2 before t_c is omitted.

If we assume that sweeps originate from single new mutations, we can use the standard expression for the rate of substitution of mutations (20) to write $\nu_a = up_a\gamma_a$ and $\nu_u = up_u\gamma_a$ where p_a is the proportion of new nonsynonymous mutations that are positively selected, and u is the mutation rate per basepair. Using Eqs. S17, we can then estimate p_a and p_u from the relations:

$$p_a = \alpha_a KA / (K_S \gamma_a) \qquad \text{(S19a)}$$

$$p_u = \alpha_u K_U / (K_S \gamma_u) \qquad \text{(S19b)}$$

In order to apply Eq. S18, it is necessary to eliminate π_0, the (unknown) diversity in the absence of selection. This was done by adding the smoothed value of $-\ln(\pi_S/\pi_0)$ for the first bin in the set (given by the linear regression of $-\ln(\pi_S/\pi_0)$ on K_A) to each of the dev_j values for the other bins, so that the terms in $\ln(\pi_0)$ cancel out. The sum of squares (*SSD*) of the resulting quantities was used as a measure of goodness of fit. A 3-dimensional grid search was performed across a defined range of values of γ_u and of the intercept and slope for γ_a, in order to obtain estimates that minimise *SSD*. For the point estimates of the parameters, we normally used a grid of 9 values for each variable, with three iterations of the search over successively narrower intervals. The resulting values of γ_a and γ_u were then used in Eqs. S19 to obtain estimates of p_a and p_u, the proportions of NS and UTR mutations that are advantageous. An estimate of π_S/π_0 for a bin was then obtained by substituting the parameter estimates for the bin into the right-hand side of Eq. S16c, providing a measure of the effect of selection at linked sites on synonymous site diversity. We also examined the effect of ignoring UTRs, by conducting similar analyses in which only NS sites were considered; in this case, only a two-dimensional grid search was needed.

We obtained estimates of $\ln(\pi_0)$ by using the fact that the predicted value of $-\ln(\pi_S)$ for the first bin in the set is equal to the sum of $\ln(\pi_0)$ and the value of $\ln(\pi_S/\pi_0)$ for the first bin predicted from Eq. S16c, denoted by P_1 (this bin is expected to have little or no effect of selective sweeps at NS sites). It follows that $\ln(\pi_0)$ can be estimated as the sum of P_1 and the observed value of $\ln(\pi_S)$ for the first bin. The predicted value of $-\ln(\pi_S)$ for any bin other than the first can then be obtained by subtracting this estimate of $\ln(\pi_0)$ from the estimate of $-\ln(\pi_S/\pi_0)$ for the bin in question. The goodness of fit of the parameter estimates was assessed by calculating the Pearson correlation between observed and predicted values across bins (omitting the first bin).

5. Correcting for the effect on K_S of weak selection at synonymous sites

We use the standard Li-Bulmer model of selection on codon usage, assuming a scaled selection coefficient γ for a given gene and a mutational bias κ, given by the ratio of the mutation rate from unpreferred to preferred codons to the mutation rate u in the reverse direction (see (11), pp. 274-275 for a description of this model). We assume that the base composition of synonymous sites is at the equilibrium value, x^*, given by the Li-Bulmer equation, such that the proportion of sites with preferred codons is equal to:

$$x^* = 1/[1+\kappa\exp(-\gamma)] \qquad \text{(S20)}$$

If a value of κ is assumed, γ can be estimated from the observed proportion of preferred codons (*Fop*) in a gene or set of genes by equating x^* to *Fop*.
The equilibrium value of the rate of substitution for sites subject to such selection is given by:

$$\lambda_s^* = 2\gamma\kappa u/[1+\kappa\exp(-\gamma)][\exp(\gamma)-1] \qquad \text{(S21a)}$$

(See (11), Eq. 6.11.)

The corresponding value for neutral sites with the same base composition, taking into account mutations in both directions, is:

$$\lambda_n^* = \kappa u[1+\exp(-\gamma)]/[1+\kappa\exp(-\gamma)] \qquad \text{(S21b)}$$

The ratio of the selected to neutral substitution rates is thus:

$$R_s = 2\gamma/[1+\exp(-\gamma)][\exp(\gamma)-1] \qquad \text{(S22)}$$

Estimates of t_{div} obtained using the neutral expectation for K_S should thus be divided by R. For the *mel-yak* data, the mean correction was 0.921 and the adjusted value of t_{div} was 3.39×10^7 generations; for *mel*, the corresponding values were 0.920 and 1.32×10^7 generations. These yield estimates of the mean rates of adaptive substitutions (ν_a) for NS sites of 3.21×10^{-10} and 2.96×10^{-10} for *mel-yak* and *mel*, respectively. The corresponding rates for UTR substitutions (ν_u) were 8.16×10^{-10} and 8.78×10^{-10}.

6. Approximating the incomplete beta function

The expected change in diversity over the two possible trajectories described by Eqs. S14a and S14b, using the approximation $q_{i0} = 1/\gamma_a$, is :

$$\frac{\Delta\pi}{\pi_0} \approx -\gamma_a^{-4r_{ij}/s} S_i^2 \qquad \text{(S23a)}$$

where S_i is the incomplete beta function:

$$S_i = \int_{q_{i0}}^{1} \left(\frac{[1-q_{it}]}{q_{it}} \right)^{2r_{ij}/s} \mathrm{d}q_{it} \qquad \text{(S23b)}$$

This leads to the following generalisation of Eq. S14d:

$$P_{ca\,j} \approx \upsilon_a \sum_i \gamma_a^{-4r_{ij}/s} S_i^2 \qquad \text{(S24)}$$

In the main text, we presented results that assume that S_i can be replaced with 1 without great loss of accuracy. We investigated the validity of this assumption by replacing S_i with 1 in Eq. S24, using the 'standard' gene model with 5 exons of 300 basepairs separated by 4 introns 100bp in length, and comparing the results with those obtained from Eqs. S23. The results are shown in Figure S8, where S_1 in the caption refers to the exact value of the mean of the sum in Eq. S24 over all synonymous sites, and S_2 to the value when S_i =1, for a range of values of γ_a. It is clear that the agreement is very close.

References

1. Nordborg M, Charlesworth B, & Charlesworth D (1996) The effect of recombination on background selection. *Genet Res* 67:159-174.
2. Charlesworth B (2015) Causes of natural variation in fitness: evidence from studies of Drosophila populations. *Proc Natl Acad Sci USA* 12:1662-1669.
3. Mank JE, Vicoso B, Berlin S, & Charlesworth B (2009) Effective population size and the faster-X effect: empirical results and their interpretation. *Evolution* 64:663-674.
4. Haddrill PR, Loewe L, & Charlesworth B (2010) Estimating the parameters of selection on nonsynonymous mutations in *Drosophila pseudoobscura* and *D. miranda*. *Genetics* 185:1381-1396.
5. Kimura M (1979) Model of effectively neutral mutations in which selective constraint is incorporated. *Proc Natl Acad Sci USA* 76:3440-3444.
6. Welch JJ, Eyre-Walker A, & Waxman D (2008) *J Mol Evol* 67:418-426.
7. Eyre-Walker A & Keightley PD (2009) Estimating the rate of adaptive mutations in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol* 26:2097-2108.
8. Kousathanas A & Keightley PD (2013) A comparison of models to infer the distribution of fitness effects of new mutations. *Genetics* 193:1197-1208.
9. Campos JL, Halligan DL, Haddrill PR, & Charlesworth B (2014) The relationship between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. *Mol Biol Evol* 31:1010-1028.
10. Barton NH (2000) Genetic hitchhiking. *Phil Trans R Soc B* 355:1553-1562.
11. Charlesworth B & Charlesworth D (2010) *Elements of Evolutionary Genetics* (Roberts and Company, Greenwood Village, CO).
12. Maynard Smith J (1971) What use is sex? *JTB* 30:319-355.
13. Weissman DB & Barton NH (2012) Limits to the rate of adaptive substitution in sexual populations. *PLoS Genetics* 8:e1002740.

14. Elyashiv E*, et al.* (2016) A genomic map of the effects of linked selection in *Drosophila*. *PloS Genetics* 12:e1006130.
15. Sattah S, Elyashiv E, Kolodny O, Rinott Y, & Sella G (2011) Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. *PLoS Genetics* 7:e1001302.
16. Wiehe THE & Stephan W (1993) Analysis of a genetic hitchiking model and its application to DNA polymorphism data. *Mol Biol Evol* 10:842-854.
17. Kim Y & Stephan W (2000) Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics* 155:1415-1427.
18. Corbett-Detig RB, Hartl DL, & Sackton TB (2015) Natural selection constrains neutral diversity across a wide range of species. *PloS Biology* 13:e1002112.
19. Kimura M (1971) Theoretical foundations of population genetics at the molecular level. *Theor Pop Biol* 2:174-208.
20. Kimura M & Ohta T (1971) *Theoretical Aspects of Population Genetics* (Princeton University Press, Princeton, N.J.).

7. SI Tables

Table S1A. Multiple linear regression coefficients for π_S for *mel – yak*

Variable	Coefficient	*P*
Intercept	-3.28×10^{-3}	0.003782
K_S	2.81×10^{-2}	$< 2 \times 10^{-16}$
Rec.	8.39×10^{-3}	$< 2 \times 10^{-16}$
Expression	-5.22×10^{-4}	$< 2 \times 10^{-16}$
K_A	-5.02×10^{-2}	$< 2 \times 10^{-16}$
CDS length	8.36×10^{-7}	0.000304
GC content	-3.97×10^{-3}	0.000567
Fop	1.59×10^{-2}	$< 2 \times 10^{-16}$

Table S1B. Multiple linear regression coefficients for π_S for *mel*

Variable	Coefficient	*P*
Intercept	-1.40×10^{-3}	0.22856
K_S	1.08×10^{-1}	$< 2 \times 10^{-16}$
Rec.	8.50×10^{-3}	$< 2 \times 10^{-16}$
Expression	-5.86×10^{-4}	$< 2 \times 10^{-16}$
K_A	-1.95×10^{-1}	6.33×10^{-15}
CDS length	5.24×10^{-7}	0.05735
GC content	-3.97×10^{-3}	0.00264
Fop	1.56×10^{-2}	1.13×10^{-15}

K_S, synonymous divergence*;* Rec., effective rate of crossing over; Expression, gene expression level across all developmental stages of *D. melanogaster; K_A*, nonsynonymous divergence*; Fop,* frequency of optimal codons in a gene*; P, P* value of a *t*-test. See the first section of the Materials and Methods for details of the data and variables used here.

Table S2A Population genomic statistics for NS sites for each bin of K_A values for *mel-yak*

Bin	*N*	K_A	π_S	β	ω_a	ω_{na}	Rec.	CDS length	$\alpha \times K_A$	*Fop*
1	122	0.0019	0.0130	0.657	0.0047	0.0049	1.128	529.1	0.0009	0.594
2	190	0.0031	0.0153	0.565	0.0069	0.0089	1.151	453.7	0.0014	0.585
3	174	0.0044	0.0153	0.537	0.0092	0.0107	1.185	545.3	0.0021	0.569
4	203	0.0056	0.0149	0.485	0.0109	0.0143	1.172	526.6	0.0025	0.562
5	189	0.0068	0.0157	0.507	0.0157	0.0153	1.249	531.1	0.0035	0.567
6	221	0.0081	0.0161	0.433	0.0166	0.0191	1.217	647.4	0.0038	0.552
7	192	0.0093	0.0154	0.492	0.0196	0.0210	1.181	549.8	0.0045	0.544
8	205	0.0106	0.0141	0.499	0.0239	0.0207	1.115	575.1	0.0057	0.545
9	185	0.0119	0.0155	0.409	0.0237	0.0260	1.210	634.5	0.0057	0.549
10	208	0.0131	0.0161	0.481	0.0303	0.0256	1.169	578.9	0.0071	0.541
11	181	0.0144	0.0158	0.511	0.0331	0.0255	1.162	615.8	0.0081	0.534
12	208	0.0156	0.0154	0.417	0.0310	0.0343	1.157	560.4	0.0074	0.537
13	164	0.0169	0.0152	0.436	0.0357	0.0342	1.204	640.1	0.0087	0.531
14	186	0.0181	0.0151	0.437	0.0383	0.0353	1.190	645.2	0.0094	0.522
15	161	0.0193	0.0151	0.423	0.0371	0.0407	1.162	541.5	0.0092	0.535
16	174	0.0206	0.0148	0.426	0.0346	0.0460	1.169	563.9	0.0088	0.518
17	146	0.0219	0.0151	0.465	0.0496	0.0367	1.207	619.9	0.0125	0.526
18	150	0.0231	0.0133	0.408	0.0446	0.0438	1.080	591.7	0.0115	0.517
19	134	0.0243	0.0146	0.421	0.0426	0.0506	1.090	513.1	0.0111	0.525
20	145	0.0256	0.0143	0.361	0.0496	0.0484	1.113	537.4	0.0129	0.514
21	127	0.0268	0.0165	0.373	0.0574	0.0486	1.128	635.0	0.0144	0.516
22	126	0.0282	0.0124	0.325	0.0317	0.0720	1.046	592.1	0.0086	0.504
23	102	0.0293	0.0149	0.316	0.0496	0.0636	1.102	538.4	0.0130	0.523
24	119	0.0305	0.0138	0.363	0.0537	0.0613	1.203	491.4	0.0142	0.508
25	90	0.0318	0.0141	0.284	0.0461	0.0717	1.082	614.1	0.0127	0.505
26	96	0.0331	0.0150	0.291	0.0575	0.0690	1.150	508.5	0.0149	0.507
27	78	0.0344	0.0143	0.316	0.0482	0.0765	1.126	520.4	0.0130	0.504
28	66	0.0355	0.0161	0.354	0.0576	0.0701	1.215	515.4	0.0160	0.506
29	125	0.0370	0.0130	0.332	0.0638	0.0722	1.114	512.7	0.0174	0.499
30	117	0.0389	0.0129	0.240	0.0582	0.0827	1.116	471.6	0.0161	0.496
31	124	0.0409	0.0132	0.324	0.0659	0.0820	1.053	442.8	0.0183	0.497
32	109	0.0430	0.0145	0.278	0.0570	0.0920	1.119	492.5	0.0164	0.500
33	102	0.0449	0.0159	0.179	0.0474	0.1095	1.189	494.5	0.0136	0.485
34	105	0.0468	0.0146	0.302	0.0774	0.0842	1.060	450.9	0.0223	0.500
35	96	0.0489	0.0136	0.277	0.0778	0.0919	1.030	401.8	0.0226	0.487
36	151	0.0521	0.0129	0.289	0.0792	0.0939	1.052	437.9	0.0236	0.472
37	131	0.0558	0.0146	0.327	0.1015	0.0908	1.117	433.4	0.0295	0.473
38	120	0.0601	0.0122	0.345	0.0911	0.1046	1.030	475.1	0.0282	0.463
39	128	0.0639	0.0118	0.185	0.0630	0.1573	1.084	481.7	0.0181	0.480
40	102	0.0679	0.0129	0.265	0.1124	0.1117	1.100	401.5	0.0340	0.467
41	100	0.0723	0.0139	0.245	0.1144	0.1185	1.094	370.1	0.0360	0.453
42	96	0.0773	0.0131	0.232	0.1176	0.1309	1.091	434.7	0.0365	0.470
43	115	0.0839	0.0136	0.252	0.1404	0.1189	1.126	416.7	0.0457	0.464
44	98	0.0914	0.0129	0.331	0.1519	0.1190	1.170	370.1	0.0519	0.466
45	100	0.0988	0.0121	0.219	0.1336	0.1658	1.148	341.0	0.0442	0.443
46	103	0.1080	0.0122	0.249	0.1786	0.1468	1.104	379.8	0.0589	0.437
47	103	0.1202	0.0128	0.210	0.1855	0.1715	1.173	408.6	0.0626	0.444
48	102	0.1343	0.0108	0.199	0.2106	0.1774	1.169	397.5	0.0727	0.417
49	93	0.1511	0.0114	0.201	0.2363	0.2027	1.229	443.4	0.0808	0.419
50	86	0.1828	0.0114	0.176	0.2566	0.2330	1.150	363.5	0.0952	0.403

Table S2B. Population genomic statistics for NS sites for each bin of K_A values for *mel*

Bin	*N*	K_A	π_S	β	ω_a	ω_{na}	Rec.	CDS length	$\alpha \times K_A$	*Fop*
1	96	0.00085	0.0157	0.577	0.0061	0.0109	1.139	678.3	0.00031	0.543
2	104	0.00105	0.0149	0.452	0.0027	0.0178	1.167	629.2	0.00014	0.545
3	128	0.00125	0.0156	0.470	0.0069	0.0165	1.142	611.0	0.00037	0.548
4	118	0.00145	0.0157	0.494	0.0125	0.0156	1.190	533.8	0.00065	0.556
5	139	0.00165	0.0148	0.448	0.0093	0.0224	1.166	590.0	0.00049	0.550
6	133	0.00185	0.0151	0.553	0.0160	0.0200	1.228	575.2	0.00084	0.543
7	131	0.00204	0.0140	0.458	0.0151	0.0244	1.100	627.1	0.00077	0.521
8	125	0.00224	0.0144	0.424	0.0162	0.0267	1.189	552.7	0.00082	0.534
9	122	0.00245	0.0150	0.469	0.0203	0.0278	1.082	612.2	0.00104	0.530
10	121	0.00264	0.0169	0.496	0.0245	0.0246	1.231	576.3	0.00132	0.534
11	122	0.00285	0.0147	0.431	0.0242	0.0282	1.085	625.1	0.00131	0.536
12	142	0.00305	0.0141	0.367	0.0237	0.0358	1.150	662.1	0.00120	0.529
13	112	0.00326	0.0147	0.532	0.0341	0.0274	1.175	536.1	0.00181	0.527
14	125	0.00345	0.0147	0.394	0.0237	0.0423	1.268	557.9	0.00126	0.526
15	130	0.00366	0.0145	0.393	0.0294	0.0376	1.172	535.2	0.00160	0.521
16	108	0.00385	0.0149	0.483	0.0418	0.0313	1.142	527.3	0.00218	0.527
17	107	0.00405	0.0150	0.403	0.0407	0.0382	1.252	564.1	0.00208	0.529
18	104	0.00425	0.0147	0.445	0.0346	0.0391	1.128	505.5	0.00197	0.530
19	101	0.00445	0.0138	0.384	0.0422	0.0428	1.068	560.1	0.00221	0.527
20	105	0.00465	0.0165	0.384	0.0397	0.0432	1.221	652.2	0.00221	0.515
21	106	0.00485	0.0136	0.343	0.0360	0.0536	1.086	588.4	0.00195	0.527
22	97	0.00504	0.0137	0.367	0.0518	0.0469	1.139	549.7	0.00266	0.521
23	97	0.00524	0.0131	0.476	0.0542	0.0454	1.037	631.7	0.00281	0.516
24	75	0.00544	0.0157	0.430	0.0522	0.0465	1.158	530.6	0.00288	0.513
25	111	0.00570	0.0136	0.318	0.0425	0.0608	1.148	505.7	0.00233	0.505
26	116	0.00600	0.0144	0.335	0.0545	0.0605	1.122	571.2	0.00281	0.502
27	122	0.00630	0.0136	0.362	0.0544	0.0598	1.167	555.0	0.00291	0.510
28	106	0.00659	0.0145	0.266	0.0316	0.0843	1.132	511.4	0.00178	0.506
29	105	0.00690	0.0130	0.331	0.0582	0.0716	1.006	572.0	0.00304	0.496
30	97	0.00721	0.0139	0.318	0.0519	0.0777	1.120	481.9	0.00289	0.516
31	107	0.00750	0.0129	0.255	0.0513	0.0873	1.069	600.1	0.00272	0.490
32	94	0.00779	0.0142	0.374	0.0711	0.0684	1.172	475.3	0.00396	0.500
33	100	0.00812	0.0128	0.259	0.0604	0.0853	1.091	433.1	0.00334	0.512
34	99	0.00856	0.0123	0.391	0.0789	0.0718	1.181	509.1	0.00439	0.500
35	117	0.00895	0.0128	0.371	0.0856	0.0817	1.102	552.4	0.00445	0.486
36	91	0.00938	0.0132	0.201	0.0636	0.1161	0.961	582.6	0.00332	0.495
37	104	0.00989	0.0132	0.275	0.0853	0.0932	1.126	488.1	0.00456	0.487
38	95	0.01038	0.0123	0.259	0.0816	0.1024	1.104	468.5	0.00456	0.484
39	103	0.01094	0.0126	0.291	0.1005	0.0969	1.083	401.2	0.00552	0.484
40	95	0.01155	0.0136	0.262	0.0994	0.1074	1.015	440.8	0.00554	0.484
41	95	0.01226	0.0129	0.334	0.1073	0.1047	1.118	426.1	0.00617	0.477
42	110	0.01293	0.0115	0.265	0.1261	0.1103	1.088	429.6	0.00687	0.462
43	99	0.01378	0.0141	0.249	0.1124	0.1208	1.065	358.0	0.00667	0.480
44	95	0.01469	0.0126	0.281	0.1417	0.1228	1.176	414.4	0.00790	0.478
45	101	0.01577	0.0125	0.182	0.1034	0.1626	1.127	373.8	0.00598	0.454
46	93	0.01694	0.0124	0.233	0.1290	0.1567	1.121	473.8	0.00760	0.433
47	107	0.01845	0.0127	0.323	0.1910	0.1452	1.186	374.4	0.01047	0.445
48	96	0.02065	0.0120	0.246	0.1674	0.1886	1.165	343.0	0.00975	0.451
49	90	0.02324	0.0118	0.164	0.2136	0.1794	1.153	424.2	0.01239	0.428
50	101	0.02712	0.0104	0.206	0.2795	0.2198	1.155	507.9	0.01501	0.424

N: number of genes in the bin; K_A: mean nonsynonymous divergence for genes in the bin from the *mel-yak* or *mel* data; π_S, mean synonymous diversity; β , shape parameter of the distribution of fitness effects (DFE); ω_a, the rate of adaptive substitutions for nonsynonymous mutations relative to the neutral rate; ω_{na}, the rate of non-adaptive substitutions (neutral or slightly deleterious) relative to the neutral rate; Rec., mean smoothed effective crossing over rates from fits of Loess regressions to the crossing over rates along each chromosome (measured in cM per Mb) in *D. melanogaster*; CDS length, coding sequence length in number of aminoacids; $\alpha \times K_A$: the mean proportion of adaptive substitutions multiplied by non-synonymous divergence; *Fop*: mean codon usage bias, measured as the frequency of optimal codons. See the first section of the Materials and Methods for further details of the sources of these data.

Table S3A. UTR population genomic statistics for each bin of K_A values for *mel-yak*

	5' UTR						3' UTR					
Bin	*N*	K_A	K_U	length	β	α	*N*	K_A	K_U	length	β	α
1	102	0.0019	0.0830	263	0.469	0.662	102	0.0019	0.0602	518	0.506	0.649
2	163	0.0031	0.0852	230	0.394	0.646	166	0.0031	0.0723	39	0.366	0.591
3	150	0.0044	0.0888	257	0.270	0.484	153	0.0044	0.0679	424	0.357	0.502
4	177	0.0056	0.0904	245	0.421	0.665	178	0.0056	0.0778	395	0.349	0.619
5	167	0.0069	0.0892	252	0.402	0.626	171	0.0068	0.0709	435	0.460	0.656
6	181	0.0081	0.0846	262	0.422	0.575	186	0.0081	0.0782	354	0.367	0.576
7	158	0.0093	0.0930	225	0.344	0.575	161	0.0093	0.0829	383	0.304	0.495
8	183	0.0106	0.1067	224	0.394	0.603	184	0.0106	0.0872	347	0.306	0.553
9	158	0.0119	0.0938	238	0.344	0.589	163	0.0119	0.0838	319	0.307	0.538
10	189	0.0131	0.0945	215	0.395	0.650	185	0.0131	0.0806	346	0.371	0.615
11	158	0.0144	0.0932	216	0.390	0.586	160	0.0144	0.0862	309	0.346	0.566
12	170	0.0156	0.0907	227	0.276	0.466	174	0.0156	0.0838	328	0.290	0.529
13	133	0.0169	0.0944	228	0.383	0.628	130	0.0169	0.0932	320	0.275	0.534
14	152	0.0181	0.0923	204	0.394	0.554	157	0.0180	0.0962	305	0.271	0.496
15	133	0.0193	0.0919	201	0.437	0.620	132	0.0193	0.1095	251	0.495	0.648
16	150	0.0205	0.1003	200	0.347	0.528	149	0.0205	0.0996	279	0.295	0.470
17	125	0.0219	0.1037	229	0.285	0.552	131	0.0219	0.0895	340	0.338	0.501
18	121	0.0231	0.1116	225	0.248	0.460	126	0.0231	0.1011	295.	0.248	0.466
19	115	0.0243	0.0984	177	0.168	0.317	111	0.0243	0.1143	226	0.401	0.655
20	118	0.0256	0.1053	161	0.370	0.574	115	0.0256	0.1192	230	0.350	0.555
21	107	0.0268	0.1095	231	0.319	0.575	112	0.0268	0.1025	293	0.333	0.628
22	105	0.0282	0.1017	199	0.357	0.575	107	0.0282	0.1129	244	0.404	0.506
23	83	0.0293	0.0991	186	0.358	0.478	85	0.0293	0.1175	233	0.206	0.411
24	98	0.0306	0.1074	171	0.261	0.438	103	0.0305	0.1140	261	0.597	0.679
25	73	0.0318	0.1022	186	0.408	0.585	72	0.0318	0.1233	230	0.436	0.656
26	74	0.0331	0.1214	206	0.333	0.488	77	0.0331	0.1182	240	0.171	0.421
27	64	0.0344	0.1072	131	0.289	0.469	59	0.0344	0.1133	204	0.303	0.576
28	50	0.0354	0.0992	186	0.398	0.490	49	0.0354	0.1156	259	0.259	0.580
29	100	0.0369	0.1073	144	0.360	0.556	100	0.0369	0.1114	198	0.216	0.473
30	93	0.0389	0.1218	166	0.434	0.713	100	0.0389	0.1239	258	0.357	0.585
31	97	0.0409	0.1152	148	0.714	0.761	98	0.0409	0.1154	204	0.436	0.586
32	91	0.0429	0.1147	165	0.376	0.578	88	0.0429	0.1357	226	0.306	0.562
33	80	0.0449	0.1129	110	0.147	0.401	84	0.0449	0.1153	166	0.369	0.634
34	82	0.0469	0.1124	148.	0.767	0.787	83	0.0469	0.1279	189	0.394	0.605
35	73	0.0489	0.1026	134	0.370	0.560	73	0.0489	0.1308	238	0.215	0.344
36	110	0.0520	0.1245	118	0.185	0.281	113	0.0520	0.1311	162	0.140	0.263
37	95	0.0558	0.1082	128	0.245	0.432	98	0.0558	0.1363	204	0.222	0.413
38	86	0.0601	0.1168	88	2.277	0.871	83	0.0601	0.1350	164	0.218	0.478
39	104	0.0639	0.1125	136	0.463	0.583	102	0.0639	0.1417	267	0.550	0.610
40	75	0.0679	0.1139	81	0.260	0.364	69	0.0680	0.1454	130	0.445	0.689
41	71	0.0723	0.1460	107	0.190	0.521	74	0.0724	0.1470	147	0.255	0.464
42	68	0.0772	0.1319	135	0.168	0.247	69	0.0772	0.1680	170	0.143	0.370
43	78	0.0841	0.1261	125	1.298	0.821	79	0.0840	0.1404	182	0.277	0.580
44	68	0.0914	0.1269	98	0.152	0.385	65	0.0914	0.1828	167	0.411	0.707
45	69	0.0988	0.1116	97	0.179	0.175	65	0.0990	0.1414	155	0.262	0.489
46	72	0.1079	0.1387	141	0.249	0.454	72	0.1081	0.1509	211	0.050	0.240
47	62	0.1202	0.1204	110	0.142	0.342	56	0.1200	0.1648	184	0.819	0.751
48	65	0.1347	0.1474	107.	0.127	0.387	63	0.1348	0.1638	183	0.248	0.542
49	57	0.1509	0.1436	83	0.447	0.500	53	0.1505	0.1670	154	0.374	0.447
50	33	0.1833	0.1422	116	0.799	0.613	34	0.1826	0.1554	142	0.259	0.196

Table S3B. UTR Population genomic statistics for each bin of K_A values for *mel*

	5' UTR						3' UTR					
Bin	*N*	K_A	K_U	length	β	α	*N*	K_A	K_U	length	β	α
1	93	0.0008	0.0199	274	0.513	0.686	94	0.0009	0.0148	408	0.475	0.607
2	92	0.0011	0.0192	289	0.402	0.535	96	0.0011	0.0172	412	0.256	0.400
3	120	0.0013	0.0180	245	0.408	0.625	122	0.0013	0.0186	399	0.419	0.645
4	107	0.0015	0.0199	253	0.334	0.606	111	0.0015	0.0172	334	0.339	0.626
5	126	0.0016	0.0177	214	0.280	0.460	129	0.0016	0.0167	380	0.383	0.622
6	126	0.0018	0.0166	248	0.312	0.479	122	0.0018	0.0188	341	0.280	0.541
7	121	0.0020	0.0168	249	0.419	0.553	121	0.0020	0.0156	362.	0.351	0.523
8	112	0.0022	0.0196	250	0.390	0.540	115	0.0022	0.0196	340	0.365	0.603
9	115	0.0024	0.0187	233	0.394	0.628	114	0.0024	0.0208	339	0.282	0.586
10	114	0.0026	0.0186	250	0.561	0.737	117	0.0026	0.0197	328	0.388	0.637
11	110	0.0029	0.0204	229	0.269	0.511	112	0.0029	0.0187	303	0.239	0.494
12	129	0.0030	0.0181	198	0.344	0.518	131	0.0030	0.0199	361	0.353	0.658
13	105	0.0033	0.0180	212	0.384	0.581	104	0.0033	0.0207	385	0.327	0.546
14	112	0.0035	0.0200	212	0.471	0.650	113	0.0035	0.0169	293	0.450	0.589
15	116	0.0037	0.0182	211	0.311	0.466	114	0.0037	0.0217	316	0.263	0.482
16	100	0.0038	0.0254	239	0.147	0.398	99	0.0038	0.0176	346	0.291	0.433
17	94	0.0041	0.0178	198	0.528	0.655	96	0.0041	0.0180	282	0.405	0.674
18	93	0.0043	0.0203	178	0.478	0.526	92	0.0043	0.0212	321	0.378	0.491
19	96	0.0045	0.0228	219	0.305	0.531	94	0.0045	0.0198	336	0.349	0.584
20	94	0.0046	0.0224	169	0.289	0.488	93	0.0046	0.0198	271.	0.348	0.561
21	94	0.0049	0.0204	223	0.434	0.579	94	0.0049	0.0207	277	0.292	0.517
22	89	0.0050	0.0177	174	0.226	0.385	94	0.0050	0.0214	230	0.367	0.595
23	86	0.0052	0.0220	195	1.244	0.806	86	0.0052	0.0215	294	0.360	0.583
24	61	0.0054	0.0177	149	0.556	0.614	61	0.0054	0.0231	230	0.336	0.608
25	104	0.0057	0.0224	172	0.463	0.658	105	0.0057	0.0232	220	0.381	0.607
26	101	0.0060	0.0218	170	0.321	0.596	105	0.0060	0.0182	243	0.643	0.655
27	106	0.0063	0.0209	201	0.544	0.709	109	0.0063	0.0220	276	0.398	0.529
28	87	0.0066	0.0228	162	0.236	0.388	91	0.0066	0.0222	226	0.295	0.524
29	91	0.0069	0.0197	156	0.498	0.581	88	0.0069	0.0339	291	0.275	0.526
30	85	0.0072	0.0238	183	0.268	0.574	86	0.0072	0.0235	209	0.263	0.494
31	93	0.0075	0.0222	186	0.195	0.433	91	0.0075	0.0231	224	0.342	0.635
32	80	0.0078	0.0206	174	0.318	0.537	82	0.0078	0.0231	216	0.233	0.454
33	90	0.0081	0.0225	178	0.229	0.446	89	0.0081	0.0209	180	0.198	0.471
34	90	0.0086	0.0243	144.	0.272	0.485	89	0.0086	0.0252	245	0.360	0.546
35	101	0.0090	0.0219	128	0.452	0.613	99	0.0090	0.0220	171	0.774	0.708
36	78	0.0094	0.0182	131	0.263	0.414	78	0.0094	0.0208	203	0.246	0.497
37	92	0.0099	0.0201	151	0.362	0.563	86	0.0099	0.0283	185	0.446	0.655
38	84	0.0104	0.0248	151	0.152	0.379	77	0.0104	0.0222	244	0.199	0.415
39	85	0.0109	0.0219	142	0.475	0.668	89	0.0109	0.0279	149	0.277	0.627
40	80	0.0116	0.0207	121	1.140	0.780	78	0.0116	0.0216	168	0.486	0.569
41	76	0.0123	0.0240	71	0.273	0.484	77	0.0123	0.0208	191	0.406	0.636
42	88	0.0129	0.0224	109	0.159	0.237	86	0.0129	0.0238	143	0.261	0.506
43	71	0.0138	0.0251	84	0.246	0.555	71	0.0138	0.0289	156	0.289	0.544
44	74	0.0147	0.0249	138	0.149	0.374	75	0.0147	0.0307	168	0.435	0.656
45	76	0.0158	0.0193	85	0.396	0.654	72	0.0157	0.0262	128	0.259	0.519
46	68	0.0170	0.0297	105	0.423	0.663	69	0.0169	0.0312	192	0.074	0.346
47	75	0.0184	0.0312	86	1.436	0.862	72	0.0184	0.0254	160	0.349	0.643
48	71	0.0207	0.0184	77	3.791	0.841	67	0.0206	0.0273	152	0.418	0.671
49	60	0.0233	0.0255	123	0.164	0.517	52	0.0234	0.0227	197	0.197	0.385
50	67	0.0270	0.0263	112	0.348	0.543	66	0.0269	0.0297	144	0.781	0.581

N: number of genes in the bin; K_A: mean nonsynonymous divergence for the genes in the bin for the *mel-yak* or *mel* data; K_U: mean UTR divergence; length: UTR length in bp; β, shape parameter of the distribution of fitness effects (DFE); α, proportion of adaptive substitutions. See the first section of the Materials and Methods for further details of the sources of these data.

Table S4 Effects of exon length on BGS effects

	Exon length		
ω_{na}	**Short**	**Standard**	**Long**
Low gene conversion rate			
0.025	1.04; 3.56	1.65; 3.55	2.53; 3.87
0.050	2.06; 4.58	3.25; 5.16	4.92; 6.26
0.075	3.01; 5.53	4.74; 6.64	7.11; 8.45
0.100	3.86; 6.38	6.02; 7.92	8.96; 10.3
0.125	4.57; 7.10	7.11; 9.02	10.4; 11.8
0.150	5.18; 7.71	7.95; 9.86	11.5; 12.9
High gene conversion rate			
0.025	0.62; 1.71	0.97; 1.83	1.49; 2.14
0.050	1.22; 2.31	1.89; 2.75	2.88; 3.52
0.075	1.75; 2.84	2.69; 3.54	4.05; 4.70
0.100	2.21; 3.30	3.36; 4.22	4.97; 5.61
0.125	2.57; 3.66	3.86; 4.71	5.61; 6.25
0.150	2.85; 3.94	4.20; 5.06	5.99; 6.64

The entries in each cell are the mean *E* values (as percentages), obtained using the summation method. The left-hand entries are the effects of NS sites alone; the right-hand entries are the net effects of NS and UTR sites. Four 100bp introns are present. The short, standard and long exons have 50, 100 and 200 codons, respectively. The mutation rate per bp is 4.5×10^{-9} and the rate of crossing over per bp is 1×10^{-8}. The low rates of gene conversion have $g_c = 1.0 \times 10^{-8}$ and $d_g = 440$; the high rates have $g_c = 5.0 \times 10^{-8}$ and $d_g = 500$. The shape parameter of the gamma distribution of fitness effects is 0.3.

Table S5. Tests of the linearity of the effect of gene length on BGS strength

ω_{na}	Predicted	Observed
Low gene conversion rate		
0.025	4.10; 4.10	3.66; 4.21
0.050	5.16; 5.80	5.33; 5.98
0.075	6.64; 7.37	6.87; 7.61
0.100	7.92; 8.76	8.20; 9.04
0.125	9.02; 9.94	9.31; 10.2
0.150	9.86; 10.9	10.2; 11.2
High gene conversion rate		
0.025	1.83; 1.98	1.89; 2.06
0.050	2.75; 2.90	2.86; 3.04
0.075	3.54; 3.71	3.69; 3.89
0.100	4.22; 4.38	4.38; 4.57
0.125	4.71; 4.89	4.87; 5.08
0.150	5.06; 5.24	5.21; 5.42

The columns labelled 'Predicted ' are the values for genes with five 300bp exons; the 'Observed' are the (unweighted) means over genes with 150, 300 and 600bp exons. The left-hand entries were obtained using the summation method with 5 exons separated by 100bp introns; the right-hand entries were obtained from the integral approximation with the mixed model of gene conversion. The other parameters are as in Table S3.

Table S6 Effects of intron length on BGS effects

	Intron length			
ω_{na}	None	Short	Standard	Long
Low gene conversion rate				
0.025	1.86; 3.93	1.75; 3.73	1.65; 3.55	1.55; 3.33
0.050	3.65; 5.72	3.42; 5.40	3.25; 5.16	3.02; 4.80
0.075	5.29; 7.37	4.99; 6.97	4.74; 6.64	4.39; 6.17
0.100	6.76; 8.83	6.33; 8.31	6.02; 7.92	5.56; 7.34
0.125	8.02; 10.1	7.50; 9.49	7.11; 9.02	6.56; 8.34
0.150	9.03; 11.1	8.41; 10.4	7.95; 9.86	7.34; 9.12
High gene conversion rate				
0.025	1.07; 1.98	1.02; 1.90	0.97; 1.83	0.92; 1.73
0.050	2.09; 3.00	1.96; 2,84	1.89; 2.75	1.77; 2.58
0.075	2.98; 3.89	2.81; 3.69	2.69; 3.54	2.54; 3.35
0.100	3.73; 4.64	3.51; 4.39	3.36; 4.22	3.15; 3.96
0.125	4.30; 5.22	4.03; 4.91	3.86; 4.71	3.61; 4.42
0.150	4.72; 5.64	4.40; 5.28	4.20; 5.06	3.93; 4.74

The entries in each cell are the mean *E* values (percentages) obtained by the summation method. The left-hand entries are the effects of NS sites alone; the right-hand entries are the net effects of NS and UTR sites.

The exon lengths are 100 codons when introns are present, and 500 codons in the absence of introns, so that the total exon length is fixed. The lengths of the short, standard and long introns are 50bp, 100bp and 200bp, respectively. The mutation rate per bp is 4.5×10^{-9} and the rate of crossing over per basepair is 1×10^{-8}. The low rates of gene conversion have $g_c = 1.0 \times 10^{-8}$ and $d_g = 350$; the high rates have $g_c = 5.0 \times 10^{-8}$ and $d_g = 500$.

Table S7. BGS effects for the *mel-yak* data with different mutation and recombination rates

	Crossing over rate		
Mutation rate	**Low**	**Standard**	**High**
Low gene conversion rate			
Low	0.18 ± 0.03	0.16 ± 0.03	0.12 ± 0.02
Standard	0.27 ± 0.05	0.24 ± 0.04	0.18 ± 0.03
High	0.37 ± 0.06	0.31 ± 0.05	0.24 ± 0.04
High gene conversion rate			
Low	0.064 ± 0.015	0.062 ± 0.013	0.086 ± 0.018
Standard	0.096 ± 0.022	0.094 ± 0.021	0.057 ± 0.012
High	0.13 ± 0.03	0.12 ± 0.03	0.11 ± 0.02

The entries in each cell are the regression coefficients of mean E for a bin on K_A, with their standard errors. The low, standard and high rates of mutation per bp are 3.0×10^{-9}, 4.5×10^{-9} and 6.0×10^{-9}, respectively. The low, standard and high rates of crossing over per bp are 0.5×10^{-8}, 1.0×10^{-8} and 2.0×10^{-8}, respectively.

Table S8. Effects of rates of mutation and crossing over on estimates of the parameters of positive selection and synonymous site diversity

Mutation rate	Crossover rate	γ_a	p_a (x 10^4)	γ_u	p_u (x 10^4)	$\pi_{r\,max}$	$\pi_{r\,mean}$	r
Low gene conversion rate								
Low	**Low**	345	1.58	151	12.3	0.900	0.768	0.911
Low	**Standard**	508	1.11	135	14.3	0.935	0.800	0.922
Low	**High**	508	1.11	119	1.61	0.963	0.854	0.918
Standard	**Low**	159	3.48	135	14.2	0.868	0.741	0.913
Standard	**Standard**	249	2.21	213	9.03	0.882	0.757	0.924
Standard	**High**	411	1.35	41.2	46.6	0.958	0.824	0.919
High	**Low**	106	5.26	151	12.3	0.823	0.701	0.914
High	**Standard**	159	3.46	213	9.03	0.850	0.727	0.925
High	**High**	220	2.42	88.1	20.2	0.939	0.811	0.921
High gene conversion rate								
Low	**Low**	508	1.11	197	9.74	0.967	0.877	0.899
Low	**Standard**	508	1.11	260	7.40	0.967	0.885	0.915
Low	**High**	508	1.11	353	5.44	0.969	0.856	0.920
Standard	**Low**	508	1.11	197	9.74	0.951	0.829	0.902
Standard	**Standard**	508	1.11	229	8.41	0.955	0.844	0.918
Standard	**High**	508	1.11	307	6.27	0.958	0.863	0.920
High	**Low**	468	1.23	338	5.69	0.907	0.775	0.908
High	**Standard**	505	1.14	370	5.21	0.917	0.787	0.920
High	**High**	508	1.11	260	7.40	0.950	0.832	0.920

The *mel-yak* data used for Figure 3 were analysed for different values of the mutation and crossing over rates, using the 'standard' gene model. The low and high rates of crossing over per bp were one-half and twice the standard value of 1 x 10^{-8}, respectively; the low and high mutation rates per bp were 3 x 10^{-9} and 6 x 10^{-9}, respectively, compared with the standard value of 4.5 x 10^{-9}. $\pi_{r\,max}$ is the maximum value of the mean synonymous site diversity of a gene relative to its value in the absence of selection; $\pi_{r\,mean}$ is the corresponding mean value over bins. Other variables are defined in the text.

Table S9. Bootstrapped *mel* estimates of parameters of positive selection and synonymous site diversity

Variable	Zero GC	Low GC	High GC
r	0.64 (0.45, 0.77)	0.65 (0.51, 0.76)	0.66 (0.54, 0.78)
Intercept for γ_a	28 (10, 110)	78 (10, 177)	350 (110, 577)
Slope for γ_a	82 (0, 167)	115 (0.0, 267)	198 (0.0, 600)
Mean γ_a	52 (29, 110)	110 (45, 210)	406 (228, 610)
p_a (x10^4)	11 (4.9, 17)	5.1 (2.6, 1.0)	0.86 (0.60, 2.1)
γ_u	123 (10, 510)	130 (10, 410)	305 (10, 610)
p_u (x10^4)	57 (3.3, 187)	34 (4.2, 187)	20 (1.0, 187)
$\pi_{r\ max}$	0.86 (0.75, 0.91)	0.89 (0.80, 0.87)	0.93 (0.89, 0.96)
$\pi_{r\ mean}$	0.79 (0.70, 0.84)	0.83 (0.75, 0.87)	0.86 (0.81, 0.90)

The data used for Figure S2 were analysed, with 250 independent bootstraps for each bin of K_A values, performed as described in the Materials and Methods. The entries show the means and (in brackets) the upper and lower 2.5 percentiles of the bootstrap distributions of the relevant parameter estimates.

8. SI Figures

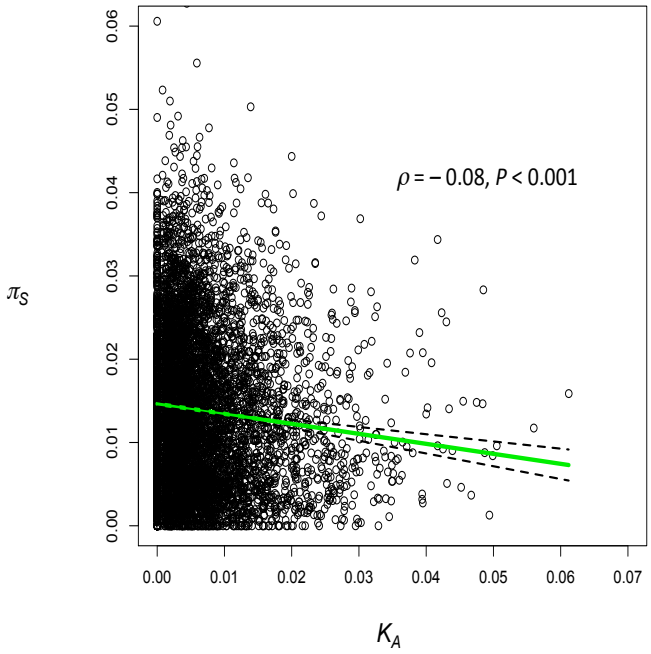

Figure S1. Synonymous site diversity (π_S) for a gene plotted against the estimated number of NS substitutions per site along the *D. melanogaster* lineage since divergence from its common ancestor with *D. simulans.* ρ is the Spearman rank correlation coefficient. The green line is the least squares linear regression; the dashed black lines represent its 95% confidence interval.

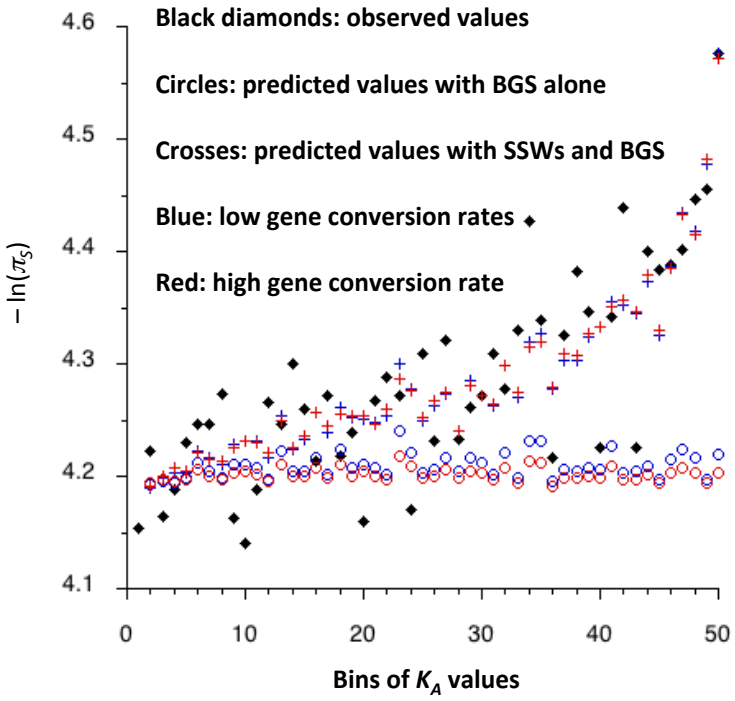

Figure S2. Plots of predicted and observed value of – ln(π_S) for each bin of K_A values for the *mel* data.

The black diamonds are the observed values of – ln(π_S) for each bin of K_A values for autosomes, corrected for the correlation between π_S and K_S as described in the first section of the Materials and Methods. The circles are the theoretical values of mean E for each bin, obtained by the integral model of BGS, assuming a single gene with 500 NS sites. The crosses are the predicted values of – ln(π_S) for each bin, given by the combined BGS and SSW models at NS and UTR sites. Red and blue correspond to the low and high gene conversion rates used in Figure 2. The mutation rate and crossing over parameters are as in Figure 2, except that large effect mutations constitute 15% of all mutations, with a selection coefficient against heterozygotes of 0.044.

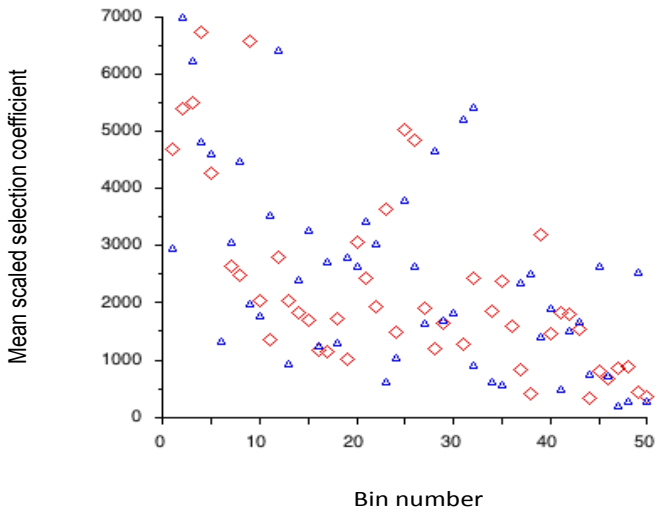

Figure S3. Mean scaled selection coefficients (γ) for NS sites for each bin of K_A values; red diamonds are *mel-yak* data, and blue triangles are *mel* data. For clarity of display, some extreme outlier values for *mel-yak* are not shown: bin 5 ($\gamma = 8.7 \times 10^3$), bin 30 ($\gamma = 1.05 \times 10^4$), and bin 32 ($\gamma = 3.04 \times 10^4$).

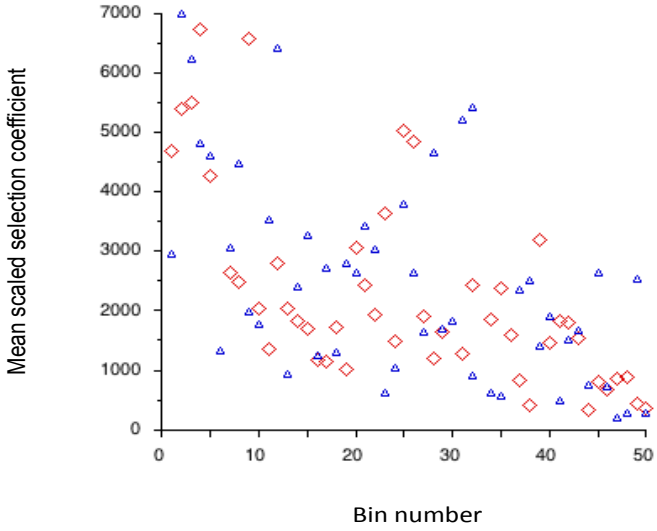

Figure S4. Mean scaled selection coefficients (γ) for UTR sites for each bin of K_A values; red diamonds are *mel-yak* data, and blue triangles are *mel* data.

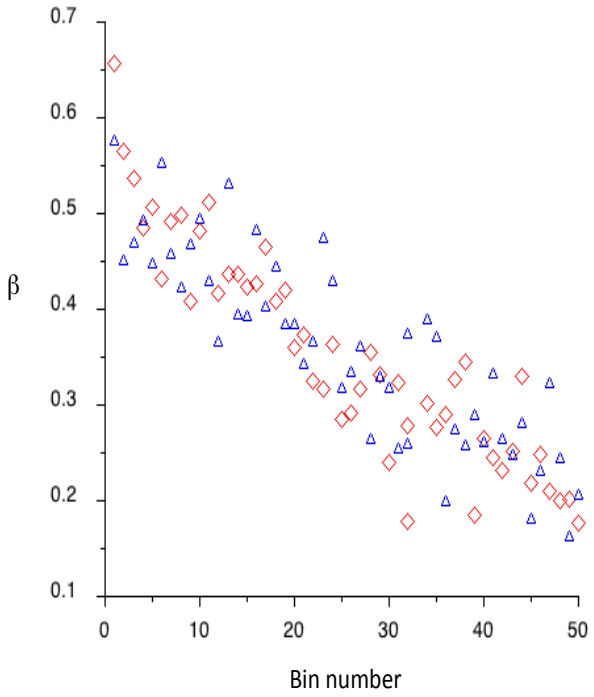

Figure S5. Shape parameters (β) for NS sites for each bin of K_A values; red diamonds are *mel-yak* data, and blue triangles are *mel* data.

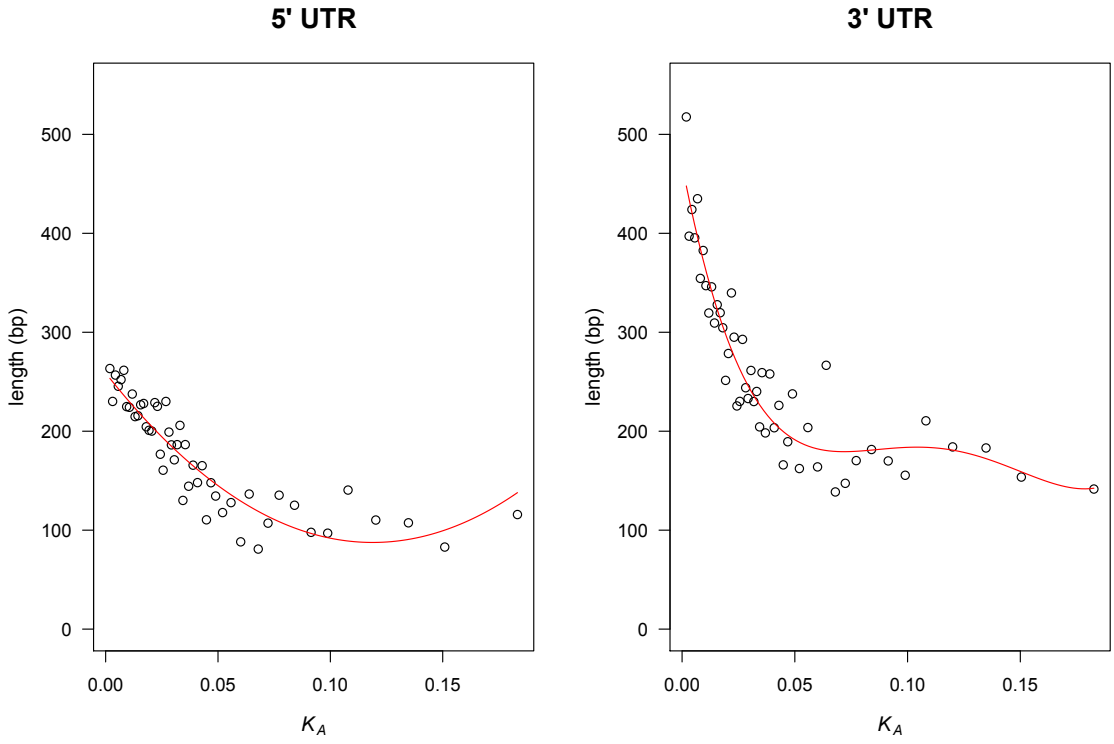

Figure S6. Plots of the lengths of the UTRs against K_A for the binned *mel-yak* data. The red curves are the quadratic least-squares fit for the 5´UTRs, and the quartic least-squares fit for the 3´UTRs.

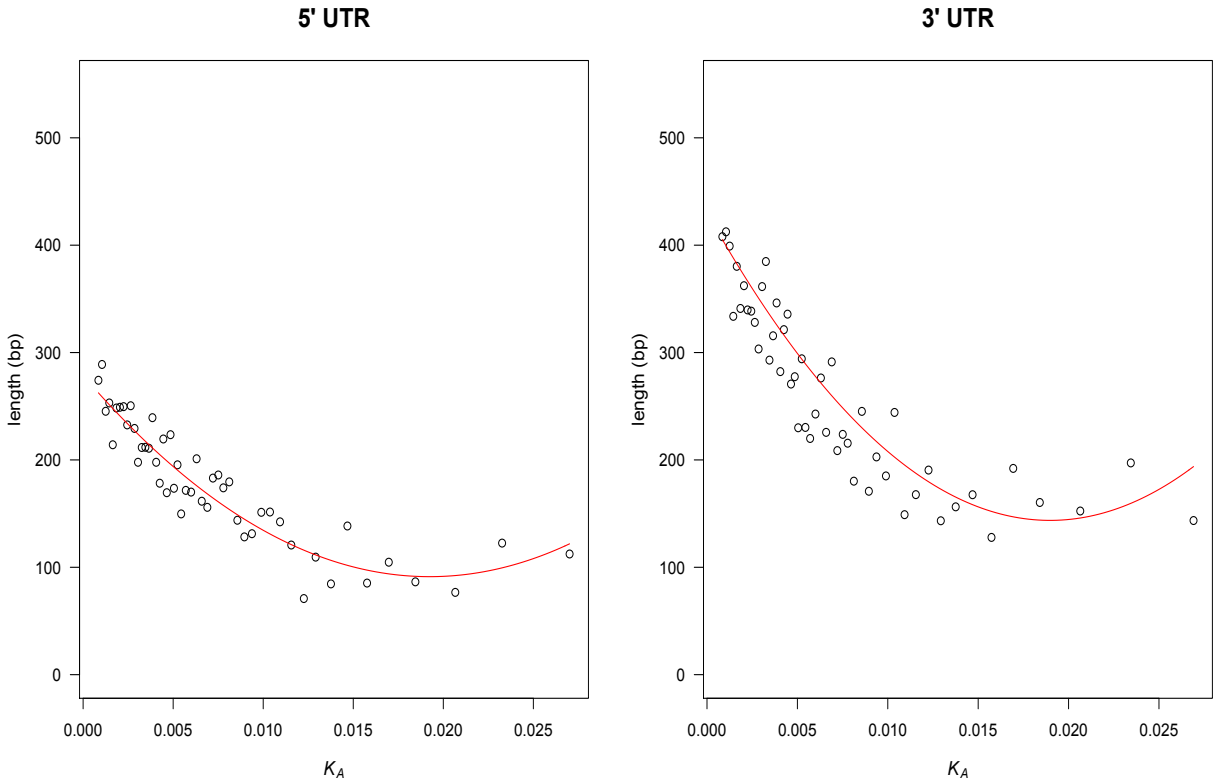

Figure S7. Plots of the lengths of the UTRs against K_A for the binned *mel* data. The red curves are the quadratic least-squares fit for the 5´UTRs and 3´UTRs.

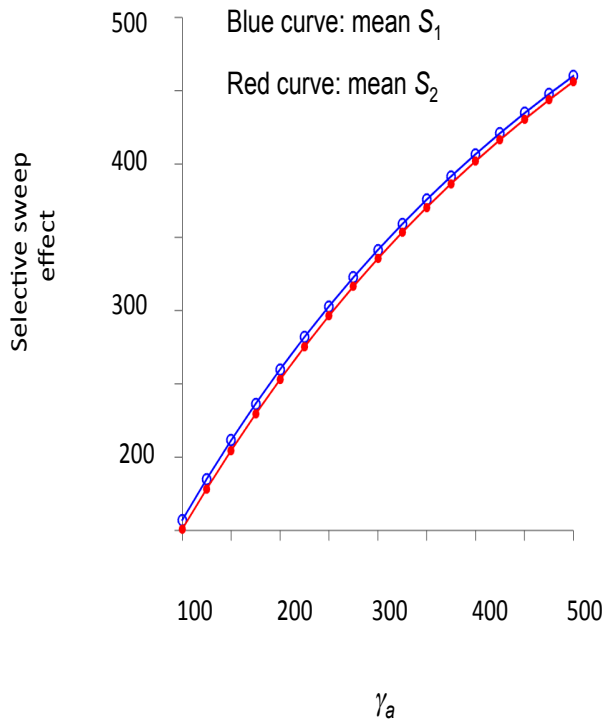

Figure S8. Plots of the mean over all synonymous sites in a gene of the predicted effects of a single selective sweep at each NS site on synonymous site diversity as a function of the scaled selection coefficient γ_a, given by the summation term in Eq. S24. S_1 is the exact result for the sum, and S_2 is the approximation with S_i =1 used in the data analyses. $N_e = 10^6$; the recombination parameters are $r_c = g_c = 10^{-8}$, and $d_g = 440$; there are 5 exons of 300 bp, separated by 4 introns of 100bp (the standard gene model).