Selective sweeps at autosomal loci with gene conversion

The expected reduction in diversity at a neutral site immediately following a sweep by a selectively favorable allele with intermediate dominance, measured relative to the initial diversity, π_0, is given by

$$\frac{\Delta\pi}{\pi_0} \approx (2N_e s)^{-\frac{4r}{s}} \qquad \text{(A1)}$$

where *s* is the selection coefficient on homozygotes for the favored allele, and *r* is the frequency of recombination between this allele and the neutral site (Charlesworth and Charlesworth 2010, p. 409).

If only gene conversion is acting, *r* is approximately twice the product of the initiation rate and the tract length of the events involved, provided that the two sites are sufficiently far apart that they are not included in the same gene conversion tract. This represents two possible types of event; a gene conversion event switching the selected allele onto an ancestral haplotype, and a gene conversion event switching a variant at the neutral site onto a haplotype carrying the selected allele. If conversion is random in direction, each of these occurs at one half of the rate of the relevant gene conversion events themselves, so that we can equate the net rate of events for an autosome to one-half the product of mean tract length and the probability of initiation of an event in female meiosis, thus correcting for the absence of events in males.

For purposes of calculation, it is convenient to rewrite equation (A1) as:

$$-\left(\frac{s}{4r}\right)\ln\left(\frac{\Delta\pi}{\pi_0}\right) \approx \ln(2N_e) + \ln(s) \qquad \text{(A2)}$$

Comeron et al. (2012) estimated the mean gene conversion tract length as 518bp and the rate of initiation of events as 1.25×10^{-7} per female meiosis, giving a value of *r* for an autosome of approximately 3.2×10^{-5} From the mean synonymous site diversity for autosomal recombining regions reported here (0.014) and the estimate of 5×10^{-9} for the per basepair mutation rate per generation in *D. melanogaster* obtained as the mean over the values for the three different strains studied by Keightley et al. (2009) and Schrider et al. (2013), we have $N_e = 7.0 \times 10^5$. Given that the estimated reduction in diversity in autosomal NC regions of *Drosophila* species in the various studies cited in the main text is 85% or more of the value for autosomal C regions, we can equate $-\Delta\pi / \pi_0$ to 0.85, so that equation (A2) becomes

$$1270s = 14.15 + \ln(s) \qquad \text{(A3)}$$

This implies that s is approximately 0.0075. Since the observed diversity includes a contribution for mutations that arose after the sweep, this represents the minimal value of s required to explain the data (even larger values would be required for soft rather than hard sweeps).

These results also assume that only a single sweep is in progress at a given time in a given NC, which is not necessarily the case (see the next section). This assumption is probably conservative, as can be seen as follows. If a second favorable mutation initiates a sweep while the first one is in progress to fixation, it will affect the final outcome with respect to level of variability only if it arises on a haplotype that lacks the first mutation. In that case, by driving this haplotype to a higher frequency than it would reach in its absence, the second favorable mutation will increase the opportunity for gene conversion events to exchange variants between haplotypes carrying the two favorable mutations, thereby increasing the final level of variability after fixation of one of the two favorable mutations over that given by equation (A3)

Can there be multiple sweeps in the autosomal NC regions?

The maximal value of ω_α for autosomal regions is approximately 0.08 (Table S4 of Supplementary Material 1); given a mutation rate per synonymous site of 5×10^{-9} (see above), this implies a substitution rate for positively selected nonsynonymous mutations of 4×10^{-10} per nonsynonymous site. The average number of codons per gene is 454 for the autosomal NC region (Campos et al. 2012); a generous estimate of the number of genes in a given NC region is 100 (Table 2). With 70% of coding sequence mutations causing an amino-acid change (Misra et al. 2002), this gives an estimate of 1.27×10^{-5} adaptive nonsynonymous substitutions per generation in an NC region, in the absence of any HRI effects. Adaptive non-coding sequence mutations may be equally important (Sella et al. 2009), so the net rate of adaptive fixations in an NC region could be as high as 2.54×10^{-5}.

Previous theoretical work has shown that beneficial mutations will establish themselves in the population without mutual interference if the mean time between the occurrence of mutations that get fixed by selection is less than the time it takes for a beneficial mutation that has established itself to spread to a high frequency or fixation, which is approximately $\ln(4N_e s_h)/s_h$ for semidominant mutations (Maynard Smith 1971; Desai and Fisher 2007), i.e. HRI due to competing beneficial mutations requires the reciprocal of the rate of substitution, K_α , of adaptive mutations in the NC regions to be smaller than the mean of $\ln(4N_e s_h)/s_h$. With the above estimate of $s = 0.0075$ to account for the observed reduction in diversity from a selective sweep with gene conversion, and with $N_e = 7.0 \times 10^5$, $\ln(4N_e s_h)/s_h$ is approximately 1.3×10^3 , which is far smaller than $1/(2.54 \times 10^{-5}) = 3.9 \times 10^4$. It is therefore very difficult to reconcile the observed rates of gene conversion and reduction in synonymous diversity in NC regions with a model in which clonal interference is solely responsible for the reduced level of adaptive evolution. This does not, of course, rule out possible joint effects of background selection and clonal interference.

Effects of weak selection on site frequency spectra

The model used is similar to that studied numerically by Zeng and Charlesworth (2009). It assumes a randomly mating, diploid, discrete-generation population with effective population size N_e. Over a long sequence of nucleotide sites, each site has two alternative types, A_1 and A_2, with mutation rates u and v from A_1 to A_2 and vice versa. (A_1 and A_2 can be taken to correspond to unpreferred versus preferred synonymous codons, respectively.) The mutational bias parameter, κ, is equal to v/u. If selection is acting, semidominance is assumed, with A_2 having a selective advantage s over A_1 when homozygous.

If the population is at statistical equilibrium, the mean numbers of sites in each state are constant over time, despite continual changes in frequencies at individual sites.

There is independence among sites, and all evolutionary forces are weak, so that diffusion approximations can be employed.

These assumptions allow the use of Wright's stationary distribution formula (Wright 1931, 1937) to describe the probability density of the frequency q of A_2 at a given site

$$\phi(q) = C\exp(\gamma q)p^{\alpha-1}q^{\beta-1} \qquad (1)$$

where $p = 1 - q$, $\alpha = 4N_e v$, $\beta = 4N_e u$, $\gamma = 2N_e s$, and the constant C is such that the integral of $\phi(q)$ between $q = 0$ and $q = 1$ is equal to 1.

When k haploid genomes are sampled from a population, the probability of obtaining a frequency x of A_2 at a given site in the sample is given by the integral of the product of $\phi(q)$ and the binomial probability of obtaining x, conditioned on q (e.g. Sawyer and Hartl 1992). This can be used to calculate quantities such as the expected value of Tajima's D for a sequence of length m (Tajima 1989) and the expected proportion of singletons in the sample, by numerical evaluation of the relevant formulae. Analyses of selection on codon usage bias in *D. melanogaster* suggest that γ is typically around 1 and κ is between 2 and 3 (Zeng and Charlesworth 2009), so that the results presented below for these values are probably most relevant for the present study.

Table S1-4. Population statistics for weakly selected sites as a function of the strength of selection γ and the mutational bias κ

γ	κ	π (%)	θ_w (%)	Tajima's D (%)	Proportion of Singletons (%)
0	1	0.980	0.977	1.14	31.1
1	1	0.908	0.920	–3.78	32.3
2	1	0.751	0.796	–16.4	35.4
4	1	0.479	0.576	–47.9	43.7
0	2	1.29	1.29	0.005	30.9
1	2	1.42	1.43	–2.88	32.0
2	2	1.33	1.40	–4.82	35.1
4	2	0.936	1.12	–48.4	43.3
0	3	1.44	1.32	0.007	30.8
1	3	1.73	1.74	–0.007	31.8
2	3	1.78	1.87	–0.593	34.7
4	3	1.37	1.63	–47.8	42.9

A sample size of 17 haploid genomes was assumed. The scaled mutation parameter β was equal to 0.02; the sequence length for calculating Tajima's *D* was 450bp.

References

Campos JL, Charlesworth B, Haddrill PR. 2012. Molecular evolution in nonrecombining regions of the *Drosophila melanogaster* genome. Genome Biol Evol. 4:278–288.

Charlesworth B, Charlesworth D. 2010. Elements of Evolutionary Genetics. Greenwood Village, Co: Roberts and Company Publishers.

Comeron JM, Ratnappan R, Bailin S. 2012. The many landscapes of recombination in *Drosophila melanogaster*. PLoS Genet. 8:e1002905.

Desai MM, Fisher DS. 2007. Beneficial mutation–selection balance and the effect of linkage on positive selection. Genetics 176:1759–1798.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177:2251–2261.

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. Genome Res. 19:1195–1201.

Maynard Smith J. 1971. What use is sex? J Theor Biol. 30:319–335.

Misra S, Crosby MA, Mungall CJ, et al. 2002. Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. Genom Biol. 3:research0083.

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. Genetics 132:1161–1176.

Schrider DR, Houle D, Lynch M, Hahn MW. 2013. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. Genetics 194:937–954.

Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the *Drosophila* genome? PLoS Genet. 5:e1000495.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Wright S. 1931. Evolution in mendelian populations. Genetics 16:97–159.

Wright S. 1937. The Distribution of Gene Frequencies in Populations. Proc Natl Acad Sci USA. 23:307–320.

Zeng K, Charlesworth B. 2009. Estimating selection intensity on synonymous codon usage in a nonequilibrium population. Genetics 183:651–662.