Supporting Information

Messer and Petrov 10.1073/pnas.1220835110

SI Text

MK Estimation of the Rate of Adaptation from Levels of Diversity and Divergence. Consider a panmictic diploid population of constant size N in a Wright–Fisher model. The expected substitution rate at a neutral site is $d_0 = 2N\mu\pi_0$, where μ is the mutation rate per generation and π_0 is the fixation probability of a neutral mutation (although $\pi_0 = 1/2N$, the notation of π_0 will be instructive). The rate of adaptive substitutions at a functional site, where new mutations may have arbitrary selection coefficients s, can be written as the difference between the overall substitution rate, minus the rate of nonadaptive substitutions:

$$d_+ = d - 2N\mu\overline{\pi} = d - d_0\frac{\overline{\pi}}{\pi_0}. \quad [S1]$$

Here, $\overline{\pi}$ specifies the average fixation probability of a nonadaptive $(s \leq 0)$ mutation at the functional site. The fraction of adaptive substitutions is therefore

$$\alpha = d_+/d = 1 - \frac{d_0}{d}\frac{\overline{\pi}}{\pi_0}. \quad [S2]$$

In practical applications, the ratio d_0/d can be inferred from sequence alignments in neutral and functional regions. Estimating the ratio $\overline{\pi}/\pi_0$, however, is typically not straightforward. One commonly used approach is to assume that most mutations in functional regions are either neutral or highly deleterious and thus restricted to very low population frequencies, whereas beneficial mutations are assumed to be rare and fix quickly (1). The polymorphism in the functional regions observed in a population sample should then primarily reflect the neutral proportion of the mutation spectrum. Under this assumption, the ratio $\overline{\pi}/\pi_0$ can be approximated by the ratio p/p_0 between the levels of polymorphism per site in the test and the neutral reference region, yielding

$$\alpha \approx 1 - \frac{d_0}{d}\frac{p}{p_0}. \quad [S3]$$

A known problem of this approach is slightly deleterious mutations. These mutations are still unlikely to become fixed in the population. They could, however, contribute noticeably to p, thereby biasing estimates of α downward. To minimize this problem, it has been proposed to exclude polymorphisms that are below a certain cutoff frequency (2, 3); the higher this cutoff, the lower the proportion of slightly deleterious polymorphisms in the sample. More sophisticated extensions of the McDonald–Kreitman (MK) test attempt to infer the actual distribution of fitness effects (DFE) of new mutations at functional sites from the site frequency spectrum (SFS) of polymorphisms at those sites, and then correct the estimates of α accordingly.

Linkage Effects on Levels of Neutral Polymorphism. It is well known that genetic draft and background selection reduce the levels of polymorphism at linked neutral sites (4, 5) and analytical formulas have been derived to estimate the predicted reduction. Specifically, when strongly deleterious mutations occur at a rate μ_d per site, background selection should reduce neutral heterozygosity H_0 by a factor $\approx \exp(-2\mu_d/r)$ (6, 7). Similarly, recurrent selective sweeps with selection coefficient s_b occurring at rate ν per site should reduce H_0 by a factor $\approx (1+8K(N)\nu s_b/r)^{-1}$, where $K(N)$ is a constant (8, 9). Under a Wright–Fisher model in a diploid population of size N and free recombination, however, we expect: $H_0 = 4N\mu_0$. Linkage effects from recurrent selective sweeps and background selection should thus reduce H_0 to

$$H_0 \approx 4N\mu_0 \times \frac{e^{-2\mu_d/r}}{1+8K(N)\nu s_b/r}. \quad [S4]$$

Linkage Effects on the SFS at Functional and Synonymous Sites. In the Wright–Fisher model under mutation–selection–drift balance and free recombination, the average number of polymorphisms with derived allele frequency x is expected to be (10, 11)

$$g(x,s) = 4N\mu_s\frac{1-e^{-4Ns(1-x)}}{(1-x)x(1-e^{-4Ns})}. \quad [S5]$$

Here, μ_s is the rate at which new mutations with selection coefficient s arise at the locus of interest per generation per individual. Integrated over the full DFE of new mutations, as specified by a density function $\rho(s)$, the expected SFS for all polymorphism at the locus is then $g(x) = \int g(x,s)\rho(s)ds$.

SI Materials and Methods

Forward Simulations of Chromosome Evolution. Our simulations model the population dynamics of a 10-Mb–long chromosome evolving in a panmictic diploid population under mutation, recombination, and selection. Genes are placed equidistantly on the chromosome with a density of one gene per 40 kb (12). Each gene consists of 8 exons of length 150 bp each, separated by introns of length 1.5 kb. Genes are flanked by a 550-bp–long 5′ UTR and a 250-bp–long 3′ UTR. We assume that three out of four sites in exons and UTRs are functional sites. Every fourth site in exons and UTRs is nonfunctional, with all mutations at those sites being neutral. These nonfunctional sites are used to model synonymous sites. Mutations occurring outside of exons or UTRs are neutral. Altogether, this yields a functional fraction of 3.75% of the chromosome. For each chromosome we store the list of mutations it harbors, with each mutation being specified by its position along the chromosome and its selection coefficient. The population consists of $N = 10^4$ diploid individuals. We assume that mutations are codominant and that fitness effects at different sites in the genome are additive. The fitness of an individual is thus given by $w = 1 + \sum_i s_i$, where the sum is taken over the selection coefficients s_i of all mutations on its two chromosomes.

Population dynamics is simulated in a model with discrete generations and constant population size. In each generation, a set of $N = 10^4$ children is newly generated. The two parents of each child are drawn from the population in the previous generation with probabilities proportional to their fitnesses. To generate the haploid gamete a parent contributes to the child, the two parental chromosomes undergo recombination at a uniform rate of $r = 10^{-8}$ per site along the chromosome (corresponding to 1 cM/Mb). Each gamete then undergoes mutation, where new mutations occur at a rate $\mu = 2.5 \times 10^{-8}$ per site per generation uniformly along the chromosome. Only the mutations which fall into exons or UTRs are followed in our simulations. Although every mutation has a specific position along the chromosome, a chromosome can harbor more than one mutation at the same site and back-mutations do not occur. Given our population parameter $N\mu = 2.5 \times 10^{-4}$, the choice of such an "effective infinite sites

model" is well justified. The simulation does not model the actual nucleotide states of mutations.

The selection coefficient of each new mutation is drawn from a specific DFE if it falls at a functional site. Mutations that fall at nonfunctional sites always have $s = 0$. After all children have been generated this way, their fitnesses are calculated and they become the parents for the next generation. At the start of a simulation run all individuals are initialized with empty chromosomes because no mutations have yet occurred. The simulations then go through a burn-in period of $10N$ generations to establish a stationary level of diversity. Every 100 generations the population is screened for fixed mutations, i.e., mutations that are present in all individuals of the population. These mutations are recorded as substitutions and removed from all chromosomes for they can no longer cause fitness differences between individuals. A simulation run is followed for 10^6 generations after the burn-in.

We estimated divergence from the mutations that became fixed during a simulation run. Polymorphism levels and frequency distributions were estimated from population samples of 100 randomly drawn chromosomes, taken every N generations throughout a run. The spectra were then averaged over all 100 samples obtained during each run. Because our chromosome has 375 kb of functional and 125 kb of synonymous sites, this corresponds to a single sample with 37.5 Mb of functional and 12.5 Mb of synonymous sites, assuming independence between samples.

The simulation is implemented in C++, making extensive use of algorithms from the GNU scientific library (13). An extended version of the simulation is implemented in the open-source program SLiM (14).

DFE-Alpha Estimation on Simulation Data. We ran DFE-alpha for each of the simulation runs specified in Table 1, using the provided online interface. These runs simulated the evolution of the above-described 10-Mb–long chromosome in a population of $N = 10^4$ diploid individuals over the course of 10^6 generations under the specific selection scenario. The SFS at functional and synonymous sites were calculated from samples of 100 randomly drawn chromosomes, taken every N generations in a simulation run. The SFS obtained from each sample were then averaged over all 100 samples taken throughout each run to generate the unfolded spectra provided to DFE-alpha. Because our 10-Mb–long chromosome has 375 kb of functional and 125 kb of synonymous sites, this corresponds to a single sample with 37.5 Mb of functional and 12.5 Mb of synonymous sites, assuming independence between samples. Divergence counts at functional and synonymous sites were inferred from the observed substitutions in each simulation run.

Asymptotic MK Estimation in Humans and Flies. Human polymorphism and divergence data are based on the resequencing of 11,404 protein coding-genes in 20 European-American individuals and were obtained from table S2 in ref. 15. A detailed description of the sequencing is provided in ref. 16. Polymorphism data for *Drosophila melanogaster* was obtained from the genome sequences of 162 inbred lines derived from Raleigh, NC (17). Only coding regions with sequence information for at least 130 strains and one-to-one orthologs across the 12 *Drosophila* species tree (18) were considered in our analysis. Each SNP was down-sampled to 130 strains, and SNPs that were no longer polymorphic after the down-sampling were removed. Divergence data with *Drosophila simulans* was obtained from probabilistic alignment kit (PRANK) alignments of the 12 *Drosophila* species. Ancestral SNP states were determined via parsimony to *D. simulans*. Functional annotation was obtained from Flybase release 5.33 (19).

1. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351(6328):652–654.
2. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. *Genetics* 158(3):1227–1234.
3. Charlesworth J, Eyre-Walker A (2008) The McDonald-Kreitman test and slightly deleterious mutations. *Mol Biol Evol* 25(6):1007–1015.
4. Stephan W (2010) Genetic hitchhiking versus background selection: The controversy and its implications. *Philos Trans R Soc Lond B Biol Sci* 365(1544):1245–1253.
5. Charlesworth B (2012) The effects of deleterious mutations on evolution at linked sites. *Genetics* 190(1):5–22.
6. Hudson RR, Kaplan NL (1995) Deleterious background selection with recombination. *Genetics* 141(4):1605–1617.
7. Stephan W, Charlesworth B, McVean G (1999) The effect of background selection at a single locus on weakly selected, partially linked variants. *Genet Res* 73:133–146.
8. Wiehe TH, Stephan W (1993) Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from Drosophila melanogaster. *Mol Biol Evol* 10(4):842–854.
9. Macpherson JM, Sella G, Davis JC, Petrov DA (2007) Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in Drosophila. *Genetics* 177(4):2083–2099.
10. Wright S (1938) The distribution of gene frequencies under irreversible mutation. *Proc Natl Acad Sci USA* 24(7):253–259.
11. Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. *Genetics* 132(4):1161–1176.
12. Lander ES, et al.; International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.
13. Galassi M, et al. (2009) *GNU Scientific Library: Reference Manual* (Network Theory, Bristol, UK), 3rd Ed.
14. Messer PW (2013) SLiM: Simulating evolution with selection and linkage. arXiv: 1301.3109.
15. Boyko AR, et al. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* 4(5):e1000083.
16. Bustamante CD, et al. (2005) Natural selection on protein-coding genes in the human genome. *Nature* 437(7062):1153–1157.
17. Mackay TF, et al. (2012) The Drosophila melanogaster Genetic Reference Panel. *Nature* 482(7384):173–178.
18. Clark AG, et al.; Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450(7167):203–218.
19. McQuilton P, St Pierre SE, Thurmond J; FlyBase Consortium (2012) FlyBase 101–the basics of navigating FlyBase. *Nucleic Acids Res* 40(Database issue):D706–D714.

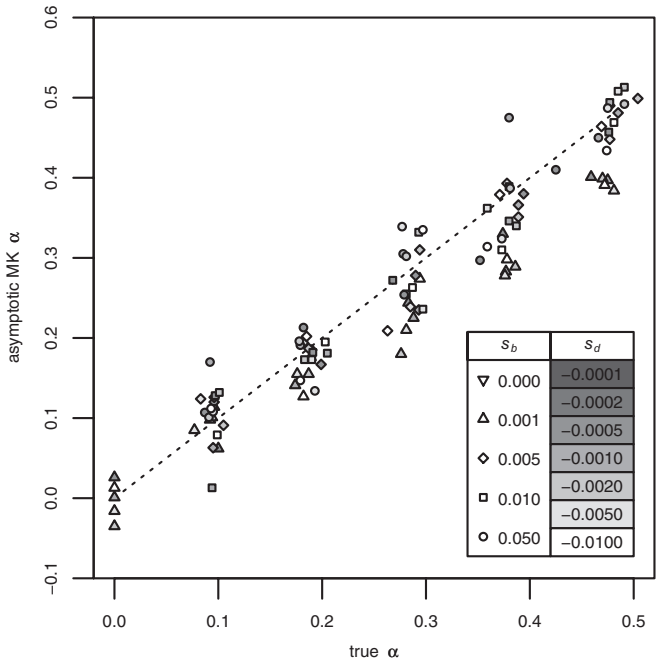

Fig. S1. Comparison of true values of α and asymptotic MK estimates for all simulation runs from Table S1. The asymptotic MK estimates were obtained by fitting $\alpha(x)$ to an exponential function of the form $\alpha(x) = a + b\exp(-cx)$ for all $x \geq 0.1$, using a nonlinear least-squares algorithm and extrapolating to $x = 1$.

Table S1. Accuracy of standard MK estimates

ρ_b^*	$s_b^\dagger$	$s_d^\ddagger$	True α	MK $\alpha^\S$	MK $\alpha^\P$
0.000000	—	−0.0001	0.00	−0.55	−0.16
0.000000	—	−0.0002	0.00	−0.27	−0.01
0.000000	—	−0.0005	0.00	−0.01	0.03
0.000000	—	−0.001	0.00	0.01	0.00
0.000000	—	−0.002	0.00	−0.01	−0.01
0.000000	—	−0.005	0.00	−0.01	−0.01
0.000000	—	−0.01	0.00	−0.02	−0.02
0.000625	0.001	−0.0001	0.08	−0.43	−0.07
0.000625	0.001	−0.0002	0.10	−0.21	0.06
0.000625	0.001	−0.0005	0.10	0.02	0.06
0.000625	0.001	−0.001	0.09	0.08	0.08
0.000625	0.001	−0.002	0.09	0.08	0.08
0.000625	0.001	−0.005	0.08	0.03	0.05
0.000625	0.001	−0.01	0.10	0.10	0.12
0.000125	0.005	−0.0001	0.09	−0.46	−0.06
0.000125	0.005	−0.0002	0.10	−0.21	0.05
0.000125	0.005	−0.0005	0.10	0.05	0.13
0.000125	0.005	−0.001	0.10	0.06	0.07
0.000125	0.005	−0.002	0.10	0.10	0.10
0.000125	0.005	−0.005	0.08	0.14	0.13
0.000125	0.005	−0.01	0.10	0.12	0.14
0.000063	0.01	−0.0001	0.09	−0.38	−0.04
0.000063	0.01	−0.0002	0.09	−0.20	0.09
0.000063	0.01	−0.0005	0.10	0.05	0.12
0.000063	0.01	−0.001	0.09	0.05	0.06
0.000063	0.01	−0.002	0.10	0.13	0.12
0.000063	0.01	−0.005	0.10	0.09	0.09
0.000063	0.01	−0.01	0.10	0.10	0.09
0.000013	0.05	−0.0001	0.07	−0.35	−0.03
0.000013	0.05	−0.0002	0.08	−0.21	0.04
0.000013	0.05	−0.0005	0.09	0.04	0.11
0.000013	0.05	−0.001	0.09	0.10	0.13
0.000013	0.05	−0.002	0.09	0.12	0.12
0.000013	0.05	−0.005	0.10	0.07	0.09
0.000013	0.05	−0.01	0.09	0.09	0.10
0.001250	0.001	−0.0001	0.17	−0.32	0.07
0.001250	0.001	−0.0002	0.19	−0.14	0.14
0.001250	0.001	−0.0005	0.19	0.11	0.17
0.001250	0.001	−0.001	0.17	0.18	0.15
0.001250	0.001	−0.002	0.19	0.13	0.12
0.001250	0.001	−0.005	0.18	0.18	0.16
0.001250	0.001	−0.01	0.18	0.19	0.15
0.000250	0.005	−0.0001	0.18	−0.34	0.04
0.000250	0.005	−0.0002	0.20	−0.14	0.18
0.000250	0.005	−0.0005	0.20	0.12	0.15
0.000250	0.005	−0.001	0.19	0.18	0.17
0.000250	0.005	−0.002	0.19	0.20	0.20
0.000250	0.005	−0.005	0.18	0.17	0.18
0.000250	0.005	−0.01	0.18	0.18	0.20
0.000125	0.01	−0.0001	0.16	−0.27	0.07
0.000125	0.01	−0.0002	0.18	−0.13	0.16
0.000125	0.01	−0.0005	0.19	0.11	0.18
0.000125	0.01	−0.001	0.21	0.17	0.19
0.000125	0.01	−0.002	0.18	0.16	0.17
0.000125	0.01	−0.005	0.19	0.21	0.19
0.000125	0.01	−0.01	0.20	0.17	0.17
0.000025	0.05	−0.0001	0.13	−0.16	0.04
0.000025	0.05	−0.0002	0.16	−0.15	0.09
0.000025	0.05	−0.0005	0.18	0.06	0.19
0.000025	0.05	−0.001	0.18	0.18	0.20
0.000025	0.05	−0.002	0.18	0.19	0.19
0.000025	0.05	−0.005	0.19	0.20	0.17
0.000025	0.05	−0.01	0.18	0.18	0.16

Table S1. Cont.

ρ_b^*	$s_b^\dagger$	$s_d^\ddagger$	True α	MK $\alpha^\S$	MK $\alpha^\P$
0.001875	0.001	−0.0001	0.26	−0.32	0.08
0.001875	0.001	−0.0002	0.28	−0.09	0.23
0.001875	0.001	−0.0005	0.28	0.20	0.23
0.001875	0.001	−0.001	0.29	0.25	0.23
0.001875	0.001	−0.002	0.28	0.24	0.21
0.001875	0.001	−0.005	0.28	0.25	0.24
0.001875	0.001	−0.01	0.29	0.27	0.26
0.000375	0.005	−0.0001	0.24	−0.26	0.12
0.000375	0.005	−0.0002	0.29	−0.09	0.21
0.000375	0.005	−0.0005	0.29	0.22	0.26
0.000375	0.005	−0.001	0.29	0.30	0.27
0.000375	0.005	−0.002	0.29	0.27	0.28
0.000375	0.005	−0.005	0.29	0.26	0.24
0.000375	0.005	−0.01	0.26	0.30	0.25
0.000188	0.01	−0.0001	0.24	−0.23	0.12
0.000188	0.01	−0.0002	0.27	−0.06	0.22
0.000188	0.01	−0.0005	0.27	0.20	0.26
0.000188	0.01	−0.001	0.28	0.28	0.27
0.000188	0.01	−0.002	0.29	0.27	0.29
0.000188	0.01	−0.005	0.30	0.28	0.27
0.000188	0.01	−0.01	0.29	0.27	0.26
0.000038	0.05	−0.0001	0.17	−0.05	0.14
0.000038	0.05	−0.0002	0.23	−0.06	0.17
0.000038	0.05	−0.0005	0.28	0.15	0.26
0.000038	0.05	−0.001	0.28	0.27	0.28
0.000038	0.05	−0.002	0.28	0.27	0.27
0.000038	0.05	−0.005	0.28	0.30	0.31
0.000038	0.05	−0.01	0.30	0.31	0.31
0.002500	0.001	−0.0001	0.35	−0.22	0.15
0.002500	0.001	−0.0002	0.40	−0.02	0.29
0.002500	0.001	−0.0005	0.37	0.29	0.30
0.002500	0.001	−0.001	0.38	0.34	0.30
0.002500	0.001	−0.002	0.39	0.33	0.29
0.002500	0.001	−0.005	0.38	0.35	0.29
0.002500	0.001	−0.01	0.38	0.36	0.32
0.000500	0.005	−0.0001	0.30	−0.18	0.21
0.000500	0.005	−0.0002	0.38	0.01	0.32
0.000500	0.005	−0.0005	0.39	0.29	0.37
0.000500	0.005	−0.001	0.39	0.38	0.36
0.000500	0.005	−0.002	0.39	0.36	0.35
0.000500	0.005	−0.005	0.38	0.37	0.39
0.000500	0.005	−0.01	0.37	0.37	0.37
0.000250	0.01	−0.0001	0.30	−0.16	0.17
0.000250	0.01	−0.0002	0.34	0.04	0.34
0.000250	0.01	−0.0005	0.38	0.29	0.38
0.000250	0.01	−0.001	0.38	0.37	0.38
0.000250	0.01	−0.002	0.39	0.38	0.35
0.000250	0.01	−0.005	0.36	0.36	0.36
0.000250	0.01	−0.01	0.37	0.38	0.33
0.000050	0.05	−0.0001	0.21	0.01	0.15
0.000050	0.05	−0.0002	0.27	0.04	0.21
0.000050	0.05	−0.0005	0.35	0.20	0.28
0.000050	0.05	−0.001	0.38	0.33	0.39
0.000050	0.05	−0.002	0.38	0.39	0.39
0.000050	0.05	−0.005	0.37	0.38	0.37
0.000050	0.05	−0.01	0.36	0.40	0.36
0.003125	0.001	−0.0001	0.41	−0.20	0.18
0.003125	0.001	−0.0002	0.48	0.04	0.36
0.003125	0.001	−0.0005	0.46	0.32	0.37
0.003125	0.001	−0.001	0.47	0.42	0.40
0.003125	0.001	−0.002	0.47	0.44	0.40
0.003125	0.001	−0.005	0.48	0.41	0.39

Table S1. Cont.

ρ_b^*	$s_b^\dagger$	$s_d^\ddagger$	True α	MK $\alpha^\S$	MK $\alpha^\P$
0.003125	0.001	−0.01	0.47	0.43	0.40
0.000625	0.005	−0.0001	0.39	−0.11	0.25
0.000625	0.005	−0.0002	0.43	0.10	0.40
0.000625	0.005	−0.0005	0.48	0.38	0.47
0.000625	0.005	−0.001	0.49	0.46	0.46
0.000625	0.005	−0.002	0.50	0.47	0.47
0.000625	0.005	−0.005	0.48	0.43	0.44
0.000625	0.005	−0.01	0.47	0.48	0.45
0.000313	0.01	−0.0001	0.33	−0.11	0.23
0.000313	0.01	−0.0002	0.42	0.08	0.36
0.000313	0.01	−0.0005	0.48	0.35	0.46
0.000313	0.01	−0.001	0.48	0.45	0.49
0.000313	0.01	−0.002	0.49	0.50	0.47
0.000313	0.01	−0.005	0.48	0.47	0.48
0.000313	0.01	−0.01	0.48	0.48	0.46
0.000063	0.05	−0.0001	0.25	0.06	0.18
0.000063	0.05	−0.0002	0.31	0.07	0.25
0.000063	0.05	−0.0005	0.42	0.28	0.41
0.000063	0.05	−0.001	0.47	0.40	0.46
0.000063	0.05	−0.002	0.49	0.47	0.47
0.000063	0.05	−0.005	0.47	0.46	0.50
0.000063	0.05	−0.01	0.47	0.48	0.46

*Fraction of adaptive mutations among all functional mutations in simulation.
†Selection coefficient of adaptive mutations in simulation.
‡Selection coefficient of deleterious mutations in simulation.
§MK estimate using cutoff $x \geq 0.1$.
¶MK estimate using cutoff $x \geq 0.5$.