Triallelic Population Genomics for Inferring Correlated Fitness Effects of Same Site Nonsynonymous Mutations

Aaron P. Ragsdale,* Alec J. Coffman,† PingHsun Hsieh,‡ Travis J. Struck,† and Ryan N. Gutenkunst†,1
*Program in Applied Mathematics, †Department of Molecular and Cellular Biology, and ‡Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721

ABSTRACT The distribution of mutational effects on fitness is central to evolutionary genetics. Typical univariate distributions, however, cannot model the effects of multiple mutations at the same site, so we introduce a model in which mutations at the same site have correlated fitness effects. To infer the strength of that correlation, we developed a diffusion approximation to the triallelic frequency spectrum, which we applied to data from *Drosophila melanogaster*. We found a moderate positive correlation between the fitness effects of nonsynonymous mutations at the same codon, suggesting that both mutation identity and location are important for determining fitness effects in proteins. We validated our approach by comparing it to biochemical mutational scanning experiments, finding strong quantitative agreement, even between different organisms. We also found that the correlation of mutational fitness effects was not affected by protein solvent exposure or structural disorder. Together, our results suggest that the correlation of fitness effects at the same site is a previously overlooked yet fundamental property of protein evolution.

KEYWORDS diffusion approximation; distribution of fitness effects; *Drosophila melanogaster*; nonsynonymous mutations; triallelic sites

MUTATIONS create genetic variation within populations, some of which causes differential fitness among individuals upon which natural selection operates. The effects of mutations on fitness range from strongly deleterious to strongly beneficial, and the distribution of fitness effects (DFE) is key for many problems in genetics, from the evolution of sex (Barton and Charlesworth 1998) to the architecture of human disease (Di Rienzo 2006). For protein-coding regions, there are generally many strongly deleterious or lethal mutations, a similar number of moderately deleterious or nearly neutral mutations, and a small number of beneficial mutations (Eyre-Walker and Keightley 2007). The DFE may be determined experimentally through direct measurements of mutation fitness effects in clonal populations of viruses, bacteria, or yeast (Wloch *et al.* 2001; Sanjuán *et al.* 2004), and recent studies have provided high-resolution DFEs for single genes (Bank *et al.* 2014; Firnberg *et al.* 2014) and for beneficial mutations (Levy *et al.* 2015). The DFE may also be inferred from comparative (Nielsen and Yang 2003; Tamuri *et al.* 2012) or population genetic (Williamson *et al.* 2005; Eyre-Walker *et al.* 2006; Keightley and Eyre-Walker 2007; Boyko *et al.* 2008) data, although these approaches have little power for strongly deleterious mutations.

In the typical population genetic approach for estimating the DFE, the population demography is first inferred using a putatively neutral class of mutations, and the DFE for another class of mutations is inferred by modeling the distribution of allele frequencies expected under a model of demography plus selection. Most population genetic inference has focused on biallelic loci, for which the ancestral allele and a single mutant (derived) allele are segregating in the population. When many individuals are sequenced, however, even single-nucleotide loci are often found to be multiallelic, with three or more segregating alleles. Multiallelic loci pose a challenge for modeling selection. To use a typical univariate DFE, one must assume that mutations at the same site all have either equal fitness effects (so that mutation location completely determines fitness) or independent fitness effects (so that mutation identity completely determines fitness). Neither of these assumptions is biologically well founded, suggesting the need for more sophisticated models of fitness

Copyright © 2016 by the Genetics Society of America
doi: 10.1534/genetics.115.184812
Manuscript received November 13, 2015; accepted for publication March 19, 2016; published Early Online March 28, 2016.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.184812/-/DC1.
[1]Corresponding author: University of Arizona, Life Sciences South Bldg., Room 325, 1007 E. Lowell St., Tucson, AZ 85271. E-mail: rgutenk@email.arizona.edu

effects. Here we introduce a model of correlated fitness effects for mutations at the same site, and we analyze sequence data to infer the strength of that correlation.

Our inference is based on triallelic codons, loci where three mutually nonsynonymous amino acid alleles are segregating in the population (Figure 1A). Interest in triallelic loci has grown recently, because such loci, while typically much less numerous than biallelic loci, are often observed in sequencing studies that sample tens or hundreds of individuals within single populations. For example, Hodgkinson and Eyre-Walker (2010) found in humans a roughly twofold excess of triallelic sites over the expectation under neutral conditions and random distribution of mutations. This led them to suggest an alternate mutational mechanism that could simultaneously generate two unique mutations, although recent population growth and substructure can account for the distribution of observed triallelic variation (Jenkins *et al.* 2014). Recently, Jenkins, Mueller, and Song (Jenkins and Song 2011; Jenkins *et al.* 2014) developed a coalescent method to calculate the expected triallelic frequency spectrum under arbitrary single-population demography. They showed that triallelic frequencies are sensitive to demographic history (Jenkins and Song 2011; Jenkins *et al.* 2014), but their method cannot model selection.

In this study, we developed a numerical diffusion simulation of expected triallelic allele frequencies for single populations with arbitrary demography and selection at one or both derived alleles. We coupled this simulation to a DFE that models the correlation between fitness effects of the two derived alleles. We applied this approach to infer the correlation coefficient of fitness effects from whole-genome *Drosophila melanogaster* data, inferring a moderate positive correlation between fitness effects of mutually nonsynonymous mutations in the same codon. To validate our inference, we compared this approach with direct biochemical experiments, finding strong agreement. Finally, we applied our approach to biologically relevant subsets of nonsynonymous mutations to assess how the fitness effects correlation varies among classes of mutations.

Theory and Methods

Here we describe the model for triallelic loci and how we solve the triallelic diffusion equation to obtain the expected sample triallelic frequency spectrum under arbitrary demography and selection. We also describe how to obtain the sample frequency spectrum under an arbitrary univariate or bivariate DFE, which we used in our inference of the correlation coefficient for selection strength at triallelic loci. Finally, we compared our results to correlation coefficients estimated from mutational scanning experiment data, discussed here as well.

Model for triallelic loci

The diffusion approximation we used is based on a triallelic extension to the standard Wright–Fisher (WF) model for allele frequency dynamics, which assumes nonoverlapping generations and random mating. The two derived alleles have selection coefficients, s_x and s_y, so their fitnesses relative to the ancestral allele are $1+s_x$ and $1+s_y$. If the two derived alleles have frequencies (i_t, j_t) in generation t in a diploid population of size N, then their frequencies in generation $t+1$ are sampled from a trinomial distribution, such that the probability of sampling (i,j) is

$$P(i,j|i_t,j_t) = \binom{2N}{i,j} p_i^i \, p_j^j \, (1-p_i-p_j)^{2N-i-j}, \qquad (1)$$

where

$$p_i = \frac{i_t(1+s_x)}{i_t(1+s_x)+j_t(1+s_y)+(2N-i_t-j_t)},$$

$$p_j = \frac{j_t(1+s_y)}{i_t(1+s_x)+j_t(1+s_y)+(2N-i_t-j_t)},$$

and $\binom{2N}{i,j}$ is the trinomial coefficient $(2N)!/\big(i!\,j!(2N-i-j)!\big)$. From here on, we focus on relative allele frequencies $x = i/2N$ and $y = j/2N$.

Most applications of the biallelic WF model assume infinite sites, so each new mutation is unique, and new mutations enter the population at a rate proportional to $\theta_{\text{bi}} = 4N_{\text{a}}\mu$. Here θ_{bi} is the population-scaled mutation rate, N_{a} is the ancestral effective population size, and μ is the per-generation mutation rate. Mutations begin at frequency $1/2N$ and are assumed to evolve independently. Given these assumptions, the density function $f(x)$ for derived allele frequencies in a population can be approximated by diffusion theory (Kimura 1964), such that the expected total number of alleles with frequency between x_0 and x_1 is $\int_{x_0}^{x_1} (\theta_{\text{bi}}/2) f(x)\, dx$, a key result from Poisson random field theory (Sawyer and Hartl 1992). The expected sample allele frequency spectrum F with n samples is then

$$F(i) = \int_0^1 \frac{\theta_{\text{bi}}}{2} f(x) \binom{n}{i} x^i (1-x)^{n-i}\, dx, \qquad (2)$$

where $\binom{n}{i}$ is the binomial coefficient. The likelihood of an observed allele frequency spectrum under this model is then a product of Poisson likelihoods for each entry in the spectrum (Sawyer and Hartl 1992).

Whereas new biallelic mutations begin at frequency $1/2N$, triallelic loci are created when a novel mutation occurs at a locus that is already biallelic. The new derived allele initially has frequency $1/2N$, and the existing derived allele has a frequency $x \in (0,1)$ drawn from the population distribution of biallelic frequencies $f(x)$ in that generation. The net rate at which triallelic loci arise is thus

$$2N\mu_{\text{tri}} \frac{\theta_{\text{bi}}}{2} f(x) = \frac{\theta_{\text{tri}}}{2} \frac{\theta_{\text{bi}}}{2} f(x), \qquad (3)$$

where μ_{tri} is the rate for mutations that hit existing biallelic sites and produce a third allele. Triallelic sites then evolve

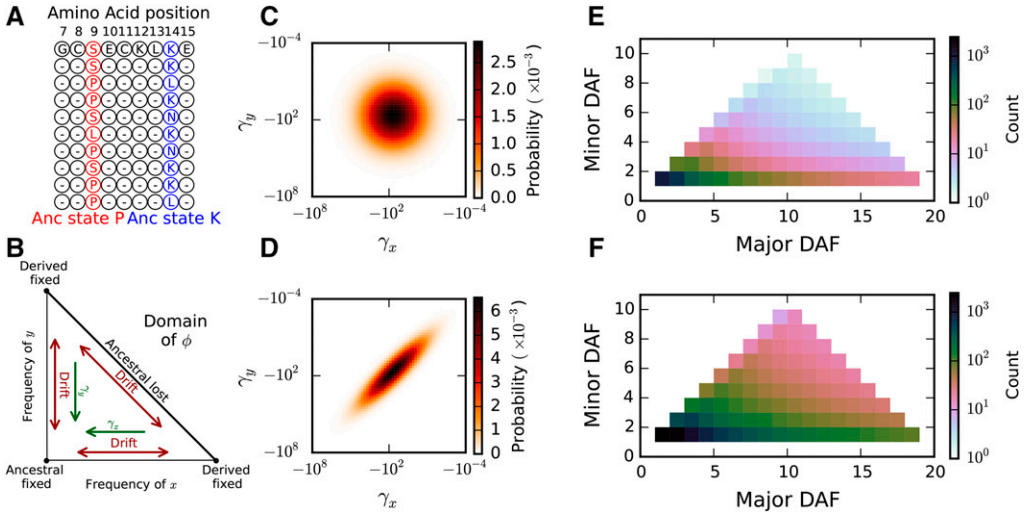

Figure 1 The triallelic frequency spectrum (TFS). (A) Mutually nonsynonymous triallelic loci in protein-coding regions have three observed segregating amino acid alleles. Here, with 10 sampled chromosomes, at position 9 the major and minor derived alleles, serine (S) and leucine (L), have frequencies 4 and 1, so this site contributes to the (4, 1) bin of the TFS. Similarly, position 14 contributes to the (2, 2) bin. (B) The domain of the triallelic diffusion equation, φ, from Equation 5. The corners correspond to fixation of one of the three alleles, and the edges correspond to loss of one of the three alleles. New mutations enter the population along the horizontal and vertical axes, with density dependent on the background biallelic frequency spectrum. Pairs of selection coefficients for the two derived nonsynonymous mutations are sampled from a bivariate DFE, which includes a parameter for correlation between selection coefficients ρ. (C) For an uncorrelated DFE, with $\rho = 0$, the selection coefficients are independent and often dissimilar. (D) For strong correlation, here $\rho = 0.9$, selection coefficients are typically very similar. (E and F) The correlation coefficient affects the expected frequency spectrum, with stronger correlation (F: $\rho = 0.9$), resulting in a higher proportion of intermediate- to high-frequency derived alleles and more triallelic sites overall relative to weak correlation (E: $\rho = 0$).

under the three-locus WF model, and we denote the density function for frequencies of triallelic loci as $\phi(x,y)$. The triallelic frequency spectrum summarizes sequence data from a sample of individuals by storing the counts of triallelic loci with each set of observed derived allele frequencies (Jenkins *et al.* 2014) (Figure 1, E and F). The expected triallelic frequency spectrum T with n samples is proportional to the integral of the density function ϕ against the trinomial sampling distribution:

$$T(i,j) \propto \int_0^1 \int_0^{1-y} \phi(x,y) \binom{n}{i,j} x^i y^j (1-x-y)^{n-i-j}\, dx\, dy. \tag{4}$$

Because the net triallelic mutation rate μ_{tri} is sensitive to mutation rate heterogeneity, in our triallelic analyses we focused on the normalized triallelic frequency spectrum, which does not depend on the overall rate of creation. Similarly, because the order in which the two derived alleles arose is often unknown, we considered only counts of major and minor derived alleles, which have respectively higher or lower sample frequencies (Figure 1). That is, for given major and minor derived allele frequencies i and j, with $j < i$, we collapsed the $T(i,j)$ and $T(j,i)$ counts together into the $T(i,j)$ bin. If in a sample we observe counts of independent triallelic frequencies $D = D(i,j)$, Poisson Random Field theory shows that the data $D(i,j)$ are Poisson distributed with mean $T(i,j)$, enabling likelihood calculations.

Diffusion approximation to the triallelic frequency spectrum with selection

To obtain the expected sample frequency spectrum for a given model of selection and demography, we numerically solved the corresponding diffusion equation. First described by Kimura (1955, 1956), the triallelic diffusion equation models the evolution of the density function $\phi(x,y)$ for the expected number of loci in the population with derived allele frequencies (x,y), such that $x,y \in (0,1)$ and $x+y<1$ (Figure 1B):

$$\frac{\partial \phi}{\partial \tau} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\frac{x(1-x)}{\nu}\phi\right) + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left(\frac{y(1-y)}{\nu}\phi\right) - \frac{\partial^2}{\partial x \partial y}\left(\frac{xy}{\nu}\phi\right) - \tilde{\gamma}_x \frac{\partial}{\partial x}\left(x(1-x)\phi\right) - \tilde{\gamma}_y \frac{\partial}{\partial y}\left(y(1-y)\phi\right). \tag{5}$$

Time τ is measured is units of $2N_a$ generations, where N_a is the ancestral effective population size. The spatial second-derivative terms account for genetic drift, which is scaled by the relative population size $\nu(\tau) = N(\tau)/N_a$, and the mixed derivative term accounts for the covariance in allele frequency changes. The population-scaled selection coefficient is $\gamma = 2N_a s$, where s is the relative fitness of the derived *vs.* ancestral allele. Here that selection coefficient must be adjusted to $\tilde{\gamma}$ to account for competition between the two segregating derived alleles, dependent on their allele frequencies. For example, if their selection coefficients are roughly equal, they will be effectively neutral when at high frequency. In general,

$$\tilde{\gamma}_x = \gamma_x \frac{1-x-y}{1-x} + (\gamma_x - \gamma_y)\frac{y}{1-x}, \tag{6}$$

with a similar expression for $\tilde{\gamma}_y$.

Like the biallelic diffusion method ∂a∂i, Equation 5 does not account for recurrent mutation, which would tend to increase derived allele frequencies. Recurrent mutation could be accounted for in the first-derivative terms, but at the cost

of additional model complexity. If it is common, neglecting recurrent mutation can bias inferences of mutation rate, population size, and selection (Desai and Plotkin 2008; Mathew *et al.* 2013). Applying our present theory thus requires that the mutation rate be high enough to create a substantial number of triallelic sites for inference, but not so high that a large fraction of biallelic or triallelic sites are affected by recurrent mutation. For most eukaryotes, including humans and *Drosophila*, mutation rates are low enough that recurrent mutation is negligible in most applications (Desai and Plotkin 2008).

Some analytic results are known for triallelic diffusion (Tier and Keller 1978; Tier 1979; Spencer and Barakat 1992), but we solved Equation 5 numerically. We used a finite-difference method similar to that in ∂a∂i (Gutenkunst *et al.* 2009). To integrate the diffusion equation forward in time, we used operator splitting to separately apply the nonmixed and mixed derivative terms each time step (Supplemental Material, File S1). We integrated the nonmixed terms, using a conservative alternating direction implicit (ADI) finite difference scheme (Chang and Cooper 1970). We integrated the mixed term, using a standard explicit scheme for mixed derivatives. We used uniform grids in x and y with equal grid spacing Δ, so that grid points lie directly on the diagonal $x + y = 1$ boundary of the domain, which readily allowed the diagonal boundary to be absorbing. Although these integration schemes worked well in the interior of the domain, application at the diagonal boundary led to an excess of density being lost (File S1 and Figure S1). To avoid this excess loss, we did not apply the ADI and mixed derivative schemes at the closest grid points to the diagonal boundary. Instead, at each time step we calculated the amount of density at each grid point that would fix along the diagonal boundary, and we directly removed that amount from the numerical density function and added it to the boundary.

To inject density into ϕ for new triallelic loci, at each time step we added density to the first interior rows of grid points based on the expected background biallelic frequency $f(x)$. For example, we added to the row of grid points $x = \Delta, 2\Delta, \ldots 1 - \Delta, y = \Delta$ with weight for point (x, Δ) proportional to the biallelic population allele density $f(x)$ at frequency x. We directly coupled with ∂a∂i to track $f(x)$. To obtain the expected sample frequency spectrum T from the population frequency spectrum ϕ, we numerically integrated against the trinomial distribution with sample size n, using Equation 4. Our code implementing these methods is integrated into ∂a∂i, available at https://bitbucket.org/gutenkunstlab/dadi.

Calculating frequency spectra under a DFE

Given a DFE, the expected sample frequency spectrum can be obtained by integrating over the expected frequency spectrum for each selection coefficient, weighted by the DFE. For biallelic sites, the DFE is a univariate distribution. For triallelic sites, the DFE is a two-dimensional joint distribution, because there are two derived alleles. Moreover, the two marginal distributions are identical, because we assume no knowledge of which allele arose first.

For our primary analysis, we used a lognormal model for the deleterious triallelic DFE (Figure 1, C and D), plus a point mass of positive selection. The lognormal distribution readily generalizes to an arbitrary number of dimensions, and the bivariate lognormal distribution includes a correlation coefficient ρ that characterizes the correlation between selection coefficients. If $\rho = 0$, the selection coefficients of the two derived alleles at a single triallelic locus are independent, whereas if $\rho = 1$, they are equal. For a fixed marginal DFE, as the correlation coefficient ρ increases, more segregating triallelic loci are expected, particularly at moderate and high derived allele frequencies (Figure 1, C–F). We quantified the relative importance of identity and location for protein mutation fitness effects through ρ; low correlation suggests that identity is more important, whereas high correlation suggests that location within the protein is more important.

To numerically integrate over the univariate DFE, we used a logarithmically spaced grid with 2000 grid points ranging from $\gamma = -2000$ to -10^{-4}, along with $\gamma = 0$ and a point mass of positive selection $\gamma_+ > 0$. Biallelic spectra were cached for each $\gamma \leq 0$, resulting in 2001 cached spectra. We assumed that alleles with $\gamma < -2000$ were effectively lethal and did not contribute to the sample frequency spectrum. We also assumed that alleles with $-10^4 < \gamma < 0$ were effectively neutral, and we used the cached spectrum for $\gamma = 0$ for contributions from this range of the DFE (Figure S2A).

To integrate over the bivariate DFE we used a logarithmically spaced grid with 50 grid points ranging from $\gamma = -2000$ to -10^{-4}, along with $\gamma = 0$ and $\gamma_+ > 0$, determined by the univariate DFE fit. We cached spectra for each possible pair (γ_x, γ_y), yielding 52^2 cached spectra. A pair of selection coefficients (γ_x, γ_y) could fall into four quadrants, depending on the sign of γ_x and γ_y. The overall frequency spectrum was calculated by summing over the weighted frequency spectra for each quadrant based on the DFE parameters p_+ and ρ. The weights were $p_+^2 + \rho\, p_+(1 - p_+)$ for both $\gamma_x, \gamma_y > 0$, $(1 - \rho)p_+(1 - p_+)$ for one selection coefficient positive and the other negative, and $(1-p_+)^2 + \rho(1 - p_+)p_+$ for both $\gamma_x, \gamma_y < 0$. These weights were found by taking the distribution of two point masses (one for positive selection, p_+, and one for negative selection, $1 - p_+$) and extending it to a bivariate distribution of point masses with correlation coefficient ρ (File S1). To integrate over the continuous distributions with one or both of the selection coefficients negative, we used the trapezoid rule. We approximated $\gamma \in (-10^{-4}, 0)$ as effectively neutral and $\gamma < -2000$ as effectively lethal (Figure S2B).

Genomic data

We extracted SNPs from phase 3 of the *Drosophila* Population Genomics Project (DPGP3) population of fruit flies from the *Drosophila* Genome Nexus Data (Lack *et al.* 2015). The data we used consist of 197 sequenced genomes from a Zambian population obtained through high-coverage haploid embryo

sequencing. This population has high genetic diversity, and it did not experience the out-of-Africa bottleneck or New World admixture that other *D. melanogaster* populations have experienced (Lack *et al.* 2015). We used Annovar (Wang *et al.* 2010) to determine the transcript and codon position of each coding SNP. The ancestral state of each codon was determined using the aligned sequences of *D. melanogaster* (April 2006, dm3) and *D. simulans* (droSim1) downloaded from the University of California, Santa Cruz genome database, by assuming that the *D. simulans* allele was ancestral. We excluded loci with no aligned *D. simulans* sequence. We downloaded the reference transcript sequences from Ensembl Biomart (Flicek *et al.* 2014) and used the ancestral states determined by the droSim1 alignment to determine the ancestral codon state.

Inferring the selection correlation coefficient

In our application to *D. melanogaster*, we used biallelic synonymous data to infer the single-population demographic history and then used nonsynonymous data to infer the parameters of the DFE. Using the unfolded synonymous allele frequency spectrum, we fitted a neutral three-epoch demographic model. This model has two instantaneous size changes, at times τ_1 and τ_2 in the past, with constant population sizes, ν_1 and ν_2, relative to the ancestral population size. We also included a parameter p_{misid} to account for ancestral state misidentification, which creates an excess of high-frequency derived alleles (Baudry and Depaulis 2003). Specifically, we compared the data not with the expected true unfolded frequency spectrum F_{true} under the demographic model, but rather with the expected observed unfolded frequency F_{obs}, such that $F_{\text{obs}}(i) = (1 - p_{\text{misid}})F_{\text{true}}(i) + p_{\text{misid}}F_{\text{true}}(n - i)$, where n is the sample size. We chose to include misidentification in our model rather than adjusting the data spectra (Hernandez *et al.* 2007), because adjusting the data leads to violations of the Poisson random field assumption, most obviously when the adjustment leads to negative entries in the data spectrum. The population-scaled mutation rate θ_{syn} was an implicit free parameter. We used the built-in optimization routines in ∂a∂i (Gutenkunst *et al.* 2009) to fit the model to the data. We fixed this demographic model for all future inferences.

The unfolded biallelic nonsynonymous allele frequency spectrum was used to infer the marginal DFE. As described above, we used a lognormal distribution for negative selection combined with a point mass of positive selection. This yielded a total of four parameters, μ and σ for the lognormal portion and γ_+ and proportion p_+ for the point mass. As in the fits for demography using synonymous data, we also included a parameter to model ancestral state misidentification. In this fit, the population-scaled mutation rate was fixed to $\theta_{\text{non}} = 2.5 \times \theta_{\text{syn}}$, and we again used ∂a∂i's optimization routines to fit the DFE to the data.

Finally, we used triallelic data with two mutually nonsynonymous derived codons to infer the correlation coefficient ρ. We fixed the demography to that inferred from the biallelic synonymous data, and we fixed the DFE parameters μ, σ, p_+, and γ_+ to the values inferred from the biallelic nonsynonymous data. This left the correlation coefficient ρ as the only free parameter of the bivariate DFE, and we also included a free parameter to account for ancestral misidentification. Assuming that the two observed derived alleles were equally likely to be the true ancestral allele, we calculated the expected observed triallelic spectrum T_{obs} from the expected true spectrum T_{true} by $T_{\text{obs}}(i,j) = (1 - p_{\text{misid}})T_{\text{true}}(i,j) + \frac{1}{2}p_{\text{misid}}T_{\text{true}}(n - i - j, j) + \frac{1}{2}p_{\text{misid}}T_{\text{true}}(i, n - i - j)$. We also left the overall population-scaled mutation rate for triallelic loci as an implicit free parameter, so our fit considered only the distribution of triallelic codons among frequency classes, not the overall number of such codons. We did this because the overall number of triallelic codons can be strongly affected by mutation rate heterogeneity, and imperfect modeling of that heterogeneity could bias our results.

We estimated model parameters by maximum composite likelihood. Following the Poisson random field framework, likelihoods $\mathcal{L}(D|\Theta)$ of the data D given the model parameters Θ were calculated by assuming that each entry in the observed triallelic frequency spectrum $D_{i,j}$ was an independent Poisson random variable with mean $T_{i,j}$ (Sawyer and Hartl 1992), where T is the expected triallelic frequency spectrum generated under Θ:

$$\mathcal{L}(\Theta|D) = \prod_{i,j} \frac{e^{-T_{i,j}} T_{i,j}^{D_{i,j}}}{D_{i,j}!}. \tag{7}$$

Because our SNP data are not actually independent, $\mathcal{L}$ is not the true likelihood, but rather a composite likelihood. To account for this, we calculated parameter uncertainties for each model fit, using the Godambe information matrix (Coffman *et al.* 2016), which adjusts the composite-likelihood statistic to account for the effects of linkage. To do so, we generated 1000 bootstrap data sets by dividing the *D. melanogaster* autosomal genome into 1000 regions of equal length and resampling among these regions.

Tests on simulated data

To generate simulated data for tests of statistical power, we first calculated the expected frequency spectrum under each model considered, using our diffusion method. To generate an observed frequency spectrum with exactly n entries, we generated n multinomial samples of frequencies, weighted by the expected frequency spectrum. To generate an observed frequency spectrum with a given mutation rate θ, we scaled the expected frequency spectrum by θ, treated the bin weights as Poisson random variables, and sampled independently for each bin.

Mutational scanning data

For comparison with our population genetic inference, we considered data from three mutational scanning studies

(Roscoe *et al.* 2013; Firnberg *et al.* 2014; Starita *et al.* 2015). Each study assayed a different protein from a different organism, using a different proxy for fitness. In all three experiments, the distribution of fitnesses was bimodal, with peaks of moderately and strongly deleterious mutations, although the relative sizes of these peaks differed markedly between experiments (Figure S3, A–C). To calculate the fitness correlation coefficient, we sampled a pair of mutually nonsynonymous mutations from each site in the protein (excluding mutations without reported fitness) and calculated the Pearson correlation of those fitnesses. The confidence intervals in Table 1 are 2.5% and 97.5% quantiles from 10,000 repetitions of this sampling. To visualize the correlations, we calculated the proportion of mutually nonsynonymous mutation pairs within each possible bin of joint fitness effects (Figure 4B and Figure S3, D–I). Because our population-genetic analysis is not sensitive to strongly deleterious mutations, we focused our analysis on moderately deleterious mutations (shaded regions in Figure S3, A–C, joint distributions in Figure S3, D–F). For details on each data set, see File S1.

Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article.

Results and Discussion

We first validated our diffusion approach to calculating the expected triallelic frequency spectrum through comparisons with coalescent simulations including demography (Figure S4) and Wright–Fisher simulations including selection (Figure S5). We then applied our method to genomic data from *D. melanogaster* to infer the strength of correlation of selection coefficients for nonsynonymous mutations that occur at the same codon in protein-coding regions. We then used simulations to characterize the performance of our approach with varying amounts of data and possible model misspecification. Finally, we compared our results to inferences made from deep mutation scanning experiments and refined our inferences to consider biologically relevant subsets of the data.

Correlation of selection strengths for nonsynonymous mutations at the same site

To estimate the correlation between fitness effects of amino acid-altering mutations, we used 197 Zambian *D. melanogaster* whole-genome sequences from the DPGP3 (Lack *et al.* 2015). We chose this population because it has high genetic diversity (and thus many triallelic sites) and a demographic history without admixture from non-sub-Saharan populations (Lack *et al.* 2015), which allowed us to model the population's demographic history using a single-population model. Recurrent mutation is expected to be rare in this population, because only $\sim$ 5% of sites are polymorphic, and of the nonsynonymous sites, only $\sim$ 4% are triallelic. As detailed in *Theory and Methods*, we first inferred demographic history using biallelic synonymous sites. We then inferred the marginal DFE for newly arising nonsynonymous mutations, using that demographic model and the biallelic nonsynonymous data. Finally, we inferred the fitness effects correlation coefficient, using our inferred demography and marginal DFE and the mutually nonsynonymous triallelic loci in the data. For all model fits, we included a parameter to account for ancestral state misidentification, which creates an excess of high-frequency derived alleles (Baudry and Depaulis 2003).

We used ∂a∂i (Gutenkunst *et al.* 2009) to fit a three-epoch population size model to the unfolded biallelic synonymous frequency spectrum (Figure 2, A and B, and Table S1). We fixed this demographic model for all future inferences, and we fitted a univariate DFE to the biallelic nonsynonymous data. For negatively selected sites ($\gamma < 0$), we assumed a lognormal distribution of selection coefficients with mean and variance parameters μ and σ, which has been previously shown to be a good approximation for the biallelic DFE for *D. melanogaster* (Kousathanas and Keightley 2013). Our DFE also included a point mass modeling a proportion p_+ of positively selected sites with scaled selection coefficient γ_+. Our inferred biallelic DFE (Figure 2C and Table S1) fits the data well (Figure 2A), with just under 1% of new mutations inferred to be beneficial (inferred $\gamma_+ = 39.9$). When fitting the DFE to the nonsynonymous data, the parameters for the lognormal portion (negatively selected sites) were tightly constrained, but p_+ and γ_+ were confounded and inversely correlated, as found in other studies (Sella *et al.* 2009; Schneider *et al.* 2011). Our inferred proportions of mutations in various selective regimes agreed well with prior work (Table S2).

We worked at the codon level to assess the correlation in selection coefficients for nonsynonymous mutations, so a triallelic locus could arise from two mutations at the same nucleotide or at different nucleotides in the same codon. We extended our inferred one-dimensional DFE to two dimensions, fixing the parameters μ, σ, γ_+, and p_+, so that the correlation coefficient ρ was the only free parameter of the bivariate lognormal distribution, along with a single parameter for ancestral misidentification. Fitting to 10,471 mutually nonsynonymous triallelic loci (Figure 3A), we inferred $\rho = 0.51$ (Figure 3B, Table 1, and Table S1). Selection coefficients for nonsynonymous mutations at the same codon are thus somewhat but not completely correlated, so location and identity play roughly equal roles in determining mutation fitness effects.

Effects of data quality and model misspecification

Statistical power to infer the selection correlation coefficient varies with the number of observed triallelic loci and the number of sampled individuals. Inference may also be biased by distortions in the observed frequency spectrum due to sequencing error or by misspecification of the demographic or selection model. To assess the sensitivity of our analysis to such effects, we considered both fits to simulated data and alternative fits to the *Drosophila* data.

Table 1 Fitness effect correlation coefficients for nonsynonymous mutations at the same codon, inferred from population genomic data and biochemical experiments

Approach	Data set	ρ	95% C.I.
Population genomic data	*D. melanogaster* all	0.51	0.45–0.57
Biochemical experiments	*E. coli* TEM-1 β-lactamase (Firnberg *et al.* 2014)	0.41	0.34–0.50
	Yeast ubiquitin (Roscoe *et al.* 2013)	0.34	0.20–0.56
	Human BRCA1 (Starita *et al.* 2015)	0.32	0.16–0.48
Population genomic data	20% most similar amino acids	0.72	0.58–0.85
	20% most dissimilar amino acids	0.24	0.15–0.32
	20% highest solvent accessibility	0.54	0.43–0.65
	20% lowest solvent accessibility	0.50	0.37–0.63
	Disordered residues	0.54	0.45–0.62
	Ordered residues	0.45	0.35–0.55

There were 10,471 mutually nonsynonymous triallelic codon polymorphisms in the 197 sampled genomes of the Zambian fruit fly data, which yielded a tight confidence interval for the selection correlation coefficient (Table 1). To test the power of our inference for different true values of the underlying correlation coefficient and smaller numbers of sampled individuals or triallelic loci, we fitted simulated data sets, assuming the exact demography and marginal DFE were known. As expected, inferences of the correlation coefficient were unbiased, and power increased with increasing number of observed triallelic loci (Figure S6, A–E). For a constant number of observed triallelic loci, the precision of the inference was insensitive to the number of sampled individuals (Figure S6F), suggesting that capturing rare triallelic variants is not crucial. To infer the correlation coefficient to a similar precision to that in the mutational-scanning studies, >2000 triallelic sites were needed, suggesting that our inference can be carried out only for populations with high genetic diversity. For example, in the 1000 Genomes Project Phase 3 human data (1000 Genomes Project Consortium 2015), among the 216 genomes from the Yoruba population, there were only 658 mutually nonsynonymous triallelic codons for which we were able to determine the ancestral state. Based on our fits to simulated data, we would not have power to accurately infer the correlation coefficient from these data.

Errors in sequencing may distort the observed site frequency spectrum, particularly at low frequencies. To test the sensitivity of our approach to sequencing error, we simulated data under our three-epoch demographic model and DFE, plus an additional model for sequencing error. The model assumed that each sequenced base had probability $\epsilon \ll 1$ to be incorrectly identified; that is, with probability ϵ, for each polymorphic site, an individual's true derived base was called as ancestral, or an individual's true ancestral base was called as derived (Johnson and Slatkin 2008). We then refitted parameters for all of our models to both the biallelic and triallelic data simulated under this model. We found that high error rates ($\epsilon \geq 10^{-4}$) biased our inference of the selection correlation coefficient upward (Figure S7). This is likely because, under this model, sequencing error reduces the proportion of alleles observed at low *vs.* moderate and high frequencies, and higher values of ρ similarly reduce the proportion of alleles expected at low frequency *vs.* high and moderate frequencies (Figure 1, C–F).

Sequencing errors may bias inference, but the DPGP3 *D. melanogaster* data we used are high-coverage (30–50 ×) haploid sequences (Lack *et al.* 2015), so we expect sequencing error was negligible in our inference. In particular, Lack *et al.* (2015) report error rates on the order 10^{-5} per site, below the 10^{-4} error rate that caused bias in our simulation study.

To assess the sensitivity of our inferences to the demographic model, we fitted two additional models to the *Drosophila* data, both simpler than the three-epoch model we focused on. For both models, we fitted the demographic parameters to the synonymous biallelic data, fitted the marginal DFE to the nonsynonymous biallelic data, and finally inferred ρ from the mutually nonsynonymous triallelic data, all as described previously. We first considered a two-epoch demographic model, consisting of a single instantaneous population size change at some time in the past. Using this model resulted in a noticeably poorer fit to the biallelic and triallelic data (Figure S8A and Table S3). The inferred lognormal portion of the marginal DFE was similar to that from the three-epoch model. Under the two-epoch model, however, we inferred more and stronger positive selection, likely because this compensates for the underestimation of high-frequency alleles in the two-epoch model (Figure S8B). This in turn caused the inferred correlation coefficient to be substantially lower (Table S3), likely because a lower correlation coefficient reduces the number of moderate- and high-frequency triallelic loci (Figure 1, C–F), partially compensating for the effect of increased positive selection. We then considered an equilibrium demography, assuming no population size changes. This model fitted the data very poorly (Figure S8A and Table S3), and the marginal DFE and the correlation coefficient we inferred were skewed toward neutrality and a lower ρ (Table S3), because these skews generate more rare variants to account for the deficit produced by this poor demographic model. Together, these analyses suggest that inference of the triallelic DFE is sensitive to misspecification of the demographic model.

In our primary *Drosophila* analysis, we assumed that the DFE followed a lognormal distribution, because such a distribution fits the biallelic data well and easily generalizes to two

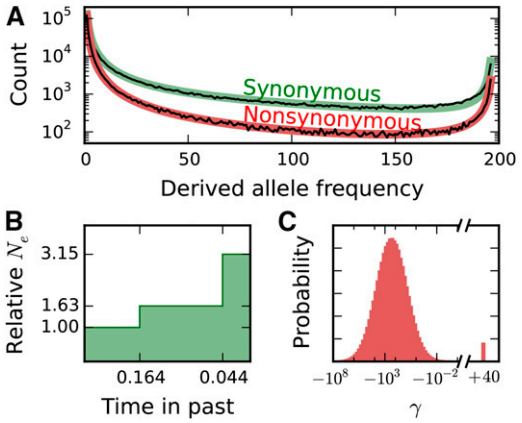

Figure 2 Inferences of demographic history and marginal distribution of fitness effects from biallelic data. (A) Biallelic synonymous and nonsynonymous data (thin black lines) and corresponding maximum-likelihood model fits (thick colored lines). Ancestral state misidentification is likely responsible for most of the excess of high-frequency derived alleles, and a parameter to model such misidentification was included in both the synonymous and nonsynonymous models. (B) Inferred demographic model, with two instantaneous population size changes. Time is in units of $2N_a$ generations, where N_a is the ancestral effective population size. (C) Inferred distribution of fitness effects, lognormally distributed for negatively selected mutations with a proportion of positively selected mutations.

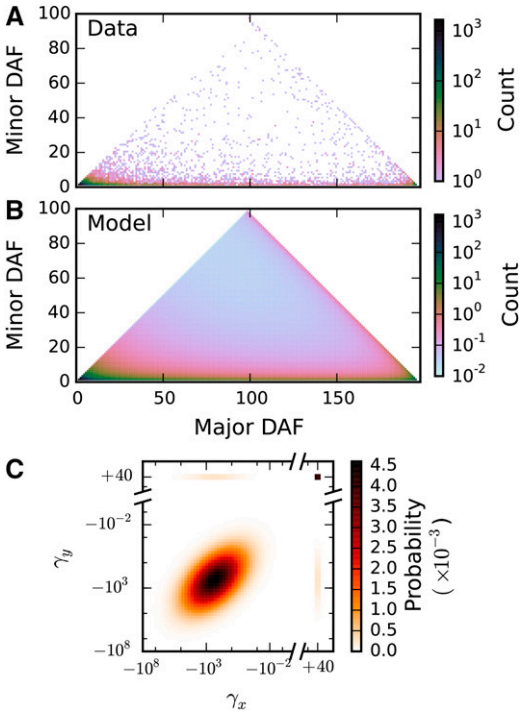

Figure 3 Inference of selection correlation coefficient from triallelic data. (A) The observed triallelic frequency spectrum for mutually nonsynonymous triallelic sites, which contained 10,471 triallelic sites. (B) The best-fit model, optimizing the correlation coefficient ρ and the ancestral misidentification parameters. (C) Joint distribution of selection coefficients from the maximum-likelihood inferred correlation coefficient of $\rho = 0.51$. Selection coefficients for nonsynonymous mutations at the same site are moderately correlated.

or more dimensions. Other analyses of the univariate DFE have, however, used other parametric distributions (Eyre-Walker *et al.* 2006; Keightley and Eyre-Walker 2007; Boyko *et al.* 2008; Kousathanas and Keightley 2013), particularly the gamma distribution (Eyre-Walker *et al.* 2006; Keightley and Eyre-Walker 2007). When we fitted a DFE with a gamma distribution for negatively selected sites and a point mass of positive selection to the bivariate data, we found a poorer fit than that of the lognormal distribution (Table S4). We nevertheless fitted a bivariate extension of the gamma distribution to the triallelic data. A number of bivariate gamma distributions have been defined (reviewed by Yue *et al.* 2001). We chose one that maintains the univariate gamma distribution when marginalized (Kibble 1941; File S1). When fitted to the *Drosophila* data, the bivariate gamma distribution yielded $\rho = 0.975$, with a moderately worse likelihood than that of the bivariate lognormal (Table S4). Note, however, that the bivariate gamma DFE is in terms of the selection coefficient γ, and the lognormal distribution is in terms of $\log\gamma$, so the correlation coefficients are not directly comparable. Given that the lognormal distribution better fits our data and has been previously found to be a good approximation for the *D. melanogaster* univariate DFE (Kousathanas and Keightley 2013), we prefer the lognormal estimate. This analysis shows, however, that the inferred correlation coefficient is sensitive to the parametric form of the bivariate distribution. Future applications may thus consider other possible forms for the bivariate DFE.

Comparison to experimental mutational scanning studies

Our population genetic approach allowed us to simultaneously study the whole genome, but it is an indirect approach to measuring the selection coefficient correlation. Complementary experimental data come from mutational scanning experiments, which use deep sequencing to simultaneously assay the function of thousands of mutant forms of a protein (Araya and Fowler 2011; Figure 4A). To measure selection coefficient correlations from such data, we sampled pairs of mutually nonsynonymous mutations for each site assayed in the protein and calculated the resulting correlations (Figure 4B and File S1). Because our population genetic inference is insensitive to strongly deleterious mutations, we restricted our analysis to the moderately deleterious mutations found in each experiment (Figure S3). We analyzed proteins from *Escherichia coli* (Firnberg *et al.* 2014), *Saccharomyces cerevisiae* (Roscoe *et al.* 2013), and humans (Starita *et al.* 2015) (Table 1). In all three cases these direct biochemical assays yielded a fitness effects correlation in agreement with our population genetic estimate, although the limited number of sites within each experiment yielded large confidence intervals, and experimental noise would tend to systematically bias the experimental correlations downward. These results suggest that the moderate correlation of mutational fitness effects we found in *D. melanogaster* also holds true for other organisms and proteins.

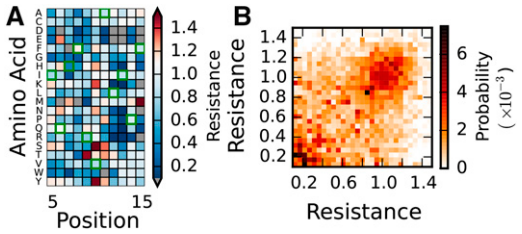

Figure 4 Mutational scanning data. (A) Partial mutational fitness landscape for *E. coli* TEM-1 β-lactamase, adapted from Firnberg *et al.* (2014). For almost all possible single mutants, fitness was assayed as relative antibiotic resistance. Gray entries denote mutations not measured, and green squares highlight the ancestral sequence. (B) Joint distribution of fitnesses for mutually nonsynonymous mutations for TEM-1 β-lactamase, using data from Firnberg *et al.* (2014). For other data sets in Table 1, see Figure S3.

Selection coefficient correlation for subsets of data

Sites within proteins vary in their evolutionary properties (Halpern and Bruno 1998; Holder *et al.* 2008), so we asked how the fitness effect correlation coefficient differs among subsets of the *D. melanogaster* population genomic data. We first tested our expectation that biochemically similar derived amino acids would have more tightly correlated selection coefficients than dissimilar derived amino acids (Yampolsky *et al.* 2005; Blanquart and Lartillot 2008). We assessed similarity, using the Grantham matrix (Grantham 1974), which scores pairs of amino acids based on similarity of biochemical properties. We then refitted the correlation coefficient and misidentification parameter to the subsets of loci with the top and bottom 20% of similarity scores. We indeed found that highly similar derived amino acids exhibited stronger correlation than dissimilar amino acids (Table 1), validating our approach.

We also assessed the correlation of fitness effects for subsets of amino acids that are buried or exposed, based on solvent accessibility, as well as subsets that are ordered or disordered, because protein structural properties are known to affect the amino acid substitution process (Dimmic *et al.* 2000). We used SPINE-D (Zhang *et al.* 2012) to separate sites into the top and bottom 20% of solvent accessibility scores and into disordered and ordered classes. For each subset, we refitted the underlying marginal DFE and then fitted the bivariate DFE to measure the correlation coefficient. As expected (Goldman *et al.* 1998; Bustamante *et al.* 2000; Tseng and Liang 2006; Lin *et al.* 2007), for buried residues with low solvent accessibility and for ordered residues, we inferred DFEs that were more negatively skewed than for residues with high solvent accessibility or that were structurally disordered (Table S5). We found, however, that these structural features did not affect the inferred fitness effects correlation coefficient (Table 1). Together, these results suggest that models of protein evolution that incorporate structural features (Wilke 2012; Arenas *et al.* 2013) do need to account for differences in the marginal DFE, but not for differences in correlation.

Conclusions

Based on the three-allele Wright–Fisher model with an influx of new mutations, we developed a novel numerical solution to the triallelic diffusion equation that simultaneously models the effects of demography and selection on pairs of derived alleles (Figure 1). Using our method, we inferred, for the first time, the correlation of mutation fitness effects at the same site within proteins from triallelic nonsynonymous SNP data (Figure 3). We found that the correlation coefficient is intermediate between completely uncorrelated and completely correlated. Early mutation–selection models of protein evolution made the unrealistic assumption that the fitness effects of multiple mutations occurring at the same site were identical (Nielsen and Yang 2003). More recent methods estimate selection coefficients for every possible amino acid at every site (Tamuri *et al.* 2012), but these complex models require a great deal of data (Tamuri *et al.* 2014). Our model of correlated fitness effects is a useful intermediate-complexity model.

We found strong quantitative agreement between the fitness effects correlation coefficient inferred from our population genomic inference and those from direct biochemical experiments (Figure 4). Moreover, this agreement held across a wide range of model organisms, for genes that vary dramatically in function, and using several measures of fitness, suggesting that this correlation of mutational fitness effects is a fundamental property of protein biology, not species or protein specific. We also refined our analysis to biologically relevant subsets of the data (Table 1). As expected, nonsynonymous pairs of similar derived amino acids show significantly higher correlation of fitness effects than dissimilar pairs. Although solvent accessibility and structural disorder did affect the marginal DFE (Table S5), we did not find a difference in fitness effects correlation among these classes of sites (Table 1). Together, our results suggest that the fitness effects correlation we inferred is a nearly universal property of protein evolution, with important implications for modeling protein evolution.

Acknowledgments

This work was supported by the National Science Foundation (DEB-1146074 to R.N.G.).

Literature Cited

Araya, C. L., and D. M. Fowler, 2011 Deep mutational scanning: assessing protein function on a massive scale. Trends Biotechnol. 29: 435–442.

Arenas, M., H. G. Dos Santos, D. Posada, and U. Bastolla, 2013 Protein evolution along phylogenetic histories under structurally constrained substitution models. Bioinformatics 29: 3020–3028.

Bank, C., R. T. Hietpas, A. Wong, D. N. Bolon, and J. D. Jensen, 2014 A Bayesian MCMC approach to assess the complete distribution of fitness effects of new mutations: uncovering the potential for adaptive walks in challenging environments. Genetics 196: 841–852.

Barton, N. H., and B. Charlesworth, 1998 Why sex and recombination? Science 281: 1986–1990.
Baudry, E., and F. Depaulis, 2003 Effect of misoriented sites on neutrality tests with outgroup. Genetics 165: 1619–1622.
Blanquart, S., and N. Lartillot, 2008 A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. 25: 842–858.
Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 4: e1000083.
Bustamante, C. D., J. P. Townsend, and D. L. Hartl, 2000 Solvent accessibility and purifying selection within proteins of Escherichia coli and Salmonella enterica. Mol. Biol. Evol. 17: 301–308.
Chang, J. S., and G. Cooper, 1970 A practical difference scheme for Fokker-Planck equations. J. Comput. Phys. 6: 1–16.
Coffman, A. J., P. Hsieh, S. Gravel, and R. N. Gutenkunst, 2016 Computationally efficient composite likelihood statistics for demographic inference. Mol. Biol. Evol. 33: 591–593.
Desai, M. M., and J. B. Plotkin, 2008 The polymorphism frequency spectrum of finitely many sites under selection. Genetics 180: 2175–2191.
Di Rienzo, A., 2006 Population genetics models of common diseases. Curr. Opin. Genet. Dev. 16: 630–636.
Dimmic, M. W., D. P. Mindell, and R. A. Goldstein, 2000 Modeling evolution at the protein level using an adjustable amino acid fitness model. Pac. Symp. Biocomput. 29: 18–29.
Eyre-Walker, A., and P. D. Keightley, 2007 The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8: 61061–61068.
Eyre-Walker, A., M. Woolfit, and T. Phelps, 2006 The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173: 891–900.
Firnberg, E., J. W. Labonte, J. J. Gray, and M. Ostermeier, 2014 A comprehensive, high-resolution map of a gene's fitness landscape. Mol. Biol. Evol. 31: 1581–1592.
Flicek, P., M. R. Amode, D. Barrell, K. Beal, K. Billis *et al.*, 2014 Ensembl 2014. Nucleic Acids Res. 42: 749–755.
Goldman, N., J. L. Thorne, and D. T. Jones, 1998 Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149: 445–458.
Grantham, R., 1974 Amino acid difference formula to help explain protein evolution. Science 185: 862–864.
Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5: e1000695.
Halpern, A. L., and W. J. Bruno, 1998 Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. 15: 910–917.
Hernandez, R. D., S. H. Williamson, and C. D. Bustamante, 2007 Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol. Biol. Evol. 24: 1792–1800.
Hodgkinson, A., and A. Eyre-Walker, 2010 Human triallelic sites: evidence for a new mutational mechanism? Genetics 184: 233–241.
Holder, M. T., D. J. Zwickl, and C. Dessimoz, 2008 Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. Philos Trans. R. Soc B 363: 4013–4021.
Jenkins, P. A., and Y. S. Song, 2011 The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. Theor. Popul. Biol. 80: 158–173.
Jenkins, P. A., J. W. Mueller, and Y. S. Song, 2014 General triallelic frequency spectrum under demographic models with variable population size. Genetics 196: 295–311.
Johnson, P. L. F., and M. Slatkin, 2008 Accounting for bias from sequencing error in population genetic estimates. Mol. Biol. Evol. 25: 199–206.
Keightley, P. D., and A. Eyre-Walker, 2007 Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177: 2251–2261.
Kibble, W. F., 1941 A two-variate gamma type distribution. Sankhya 5: 137–150.
Kimura, M., 1955 Random genetic drift in multi-allelic locus. Evolution 9: 419–435.
Kimura, M., 1956 Random genetic drift in a tri-allelic locus; exact solution with a continuous model. Biometrics 12: 57–66.
Kimura, M., 1964 Diffusion models in population genetics. J. Appl. Probab. 1: 177–232.
Kousathanas, A., and P. D. Keightley, 2013 A comparison of models to infer the distribution of fitness effects of new mutations. Genetics 193: 1197–1208.
Lack, J. B., C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig *et al.*, 2015 The *Drosophila* Genome Nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. Genetics 199: 1229–1241.
Levy, S. F., J. R. Blundell, S. Venkataram, D. A. Petrov, D. S. Fisher *et al.*, 2015 Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature 519: 181–186.
Lin, Y. S., W. L. Hsu, J. K. Hwang, and W. H. Li, 2007 Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. Mol. Biol. Evol. 24: 1005–1011.
Mathew, L. A., P. R. Staab, L. E. Rose, and D. Metzler, 2013 Why to account for finite sites in population genetic studies and how to do this with Jaatha 2.0. Ecol. Evol. 3: 3647–3662.
Nielsen, R., and Z. Yang, 2003 Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol. Biol. Evol. 20: 1231–1239.
1000 Genomes Project Consortium, 2015 A global reference for human genetic variation. Nature 526: 68–74.
Roscoe, B. P., K. M. Thayer, K. B. Zeldovich, D. Fushman, and D. N. A. Bolon, 2013 Analyses of the effects of all ubiquitin point mutants on yeast growth rate. J. Mol. Biol. 425: 1363–1377.
Sanjuán, R., A. Moya, and S. F. Elena, 2004 The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc. Natl. Acad. Sci. USA 101: 8396–8401.
Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics 132: 1161–1176.
Schneider, A., B. Charlesworth, A. Eyre-Walker, and P. D. Keightley, 2011 A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189: 1427–1437.
Sella, G., D. A. Petrov, M. Przeworski, and P. Andolfatto, 2009 Pervasive natural selection in the Drosophila genome? PLoS Genet. 5: e1000495.
Spencer, H. G., and R. Barakat, 1992 Random genetic drift and selection in a triallelic locus: a continuous diffusion model. Math. Biosci. 108: 127–139.
Starita, L. M., D. L. Young, M. Islam, J. O. Kitzman, J. Gullingsrud *et al.*, 2015 Massively parallel functional analysis of BRCA1 RING domain variants. Genetics 200: 413–422.
Tamuri, A. U., M. dos Reis, and R. A. Goldstein, 2012 Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. Genetics 190: 1101–1115.
Tamuri, A. U., N. Goldman, and M. dos Reis, 2014 A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. Genetics 197: 257–271.
Tier, C., 1979 A tri-allelic diffusion model with selection, migration, and mutation. Math. Biosci. 60: 41–60.

Tier, C., and J. B. Keller, 1978 A tri-allelic diffusion model with selection. SIAM J. Appl. Math. 35: 521–535.
Tseng, Y. Y., and J. Liang, 2006 Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a Bayesian Monte Carlo approach. Mol. Biol. Evol. 23: 421–436.
Wang, K., M. Li, and H. Hakonarson, 2010 ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38: 1–7.
Wilke, C. O., 2012 Bringing molecules back into molecular evolution. PLoS Comput. Biol. 8: 6–9.
Williamson, S. H., R. Hernandez, A. Fledel-alon, L. Zhu, R. Nielsen *et al.*, 2005 Simultanous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. USA 102: 7882–7887.
Wloch, D. M., K. Szafraniec, R. H. Borts, and R. Korona, 2001 Direct estimate of the mutation rate and the distribution of fitness effects in the yeast *Saccharomyces cerevisiae*. Genetics 159: 441–452.
Yampolsky, L. Y., F. A. Kondrashov, and A. S. Kondrashov, 2005 Distribution of the strength of selection against amino acid replacements in human proteins. Hum. Mol. Genet. 14: 3191–3201.
Yue, S., T. B. M. J. Ouarda, and B. Bobée, 2001 A review of bivariate gamma distributions for hydrological application. J. Hydrol. 246: 1–18.
Zhang, T., E. Faraggi, B. Xue, A. K. Dunker, V. N. Uversky *et al.*, 2012 SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. J. Biomol. Struct. Dyn. 29: 799–813.

Communicating editor: Y. S. Song

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.184812/-/DC1

Triallelic Population Genomics for Inferring Correlated Fitness Effects of Same Site Nonsynonymous Mutations

Aaron P. Ragsdale, Alec J. Coffman, PingHsun Hsieh, Travis J. Struck, and Ryan N. Gutenkunst

Copyright © 2016 by the Genetics Society of America
DOI: 10.1534/genetics.115.184812

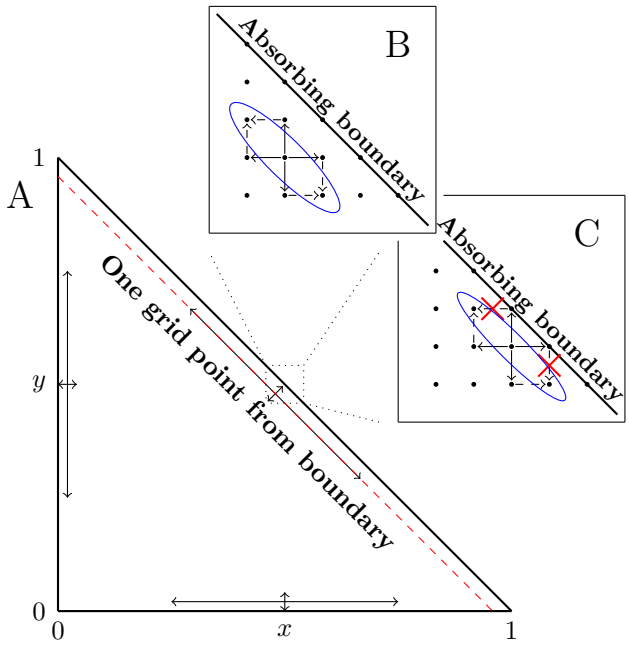

Figure S1: **Numerical challenges along diagonal boundary.** A: In the interior of the domain, we integrated forward with Eq. S1 and Eq. S2. B: Near the boundary of the domain, density spreads primarily diagonally (blue ellipse). In our scheme, the ADI step (Eq. S1, solid arrows) pushes density horizontally and vertically, and the explicit step (Eq. S2, dashed arrows) accounts for diagonal flow. C: Immediately adjacent to the boundary, the ADI step pushes excess density onto the boundary, which cannot return (red Xs), because the boundary is absorbing. We thus applied the ADI step to all points except those adjacent to the boundary. At the adjacent points we calculated the density that should be lost to the boundary each timestep and explicitly moved that much density to the boundary.

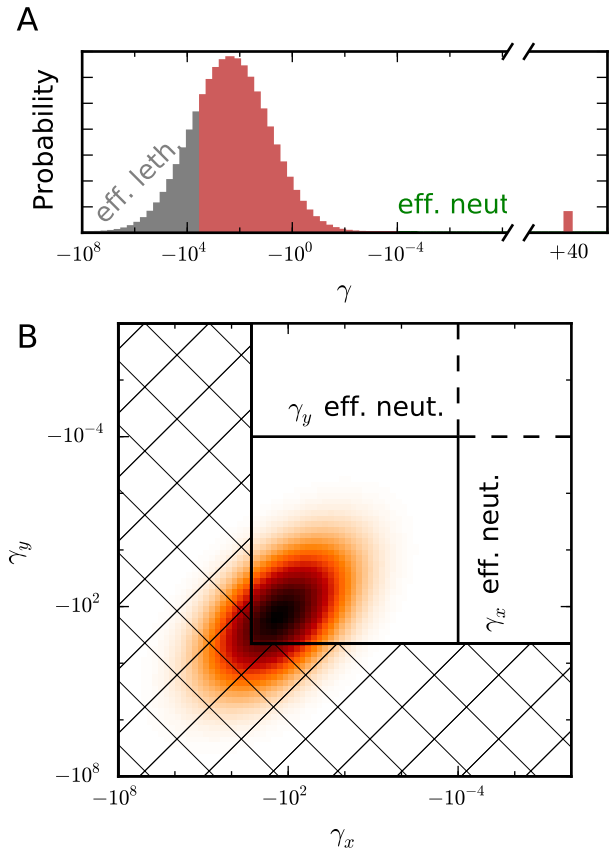

Figure S2: **Integration over DFEs.** A: Best-fit lognormal DFE for bivariate nonsynonymous data, with regimes assumed to be effectively neutral and lethal highlighted. B: Best-fit bivariate lognormal DFE (not showing positive selection portion) for all triallelic nonsynonymous data. In the crosshatched area, we considered one or both alleles to be effectively lethal, such that they did not contribute to observed segregating variation. In regions for which γ_x or γ_y were assumed effectively neutral, we used $\gamma = 0$ for calculation. For example, if $\gamma_x = -10^{-6}$ and $\gamma_y = -1$, we used the frequency spectrum calculated for $(\gamma_x, \gamma_y) = (0, -1)$.

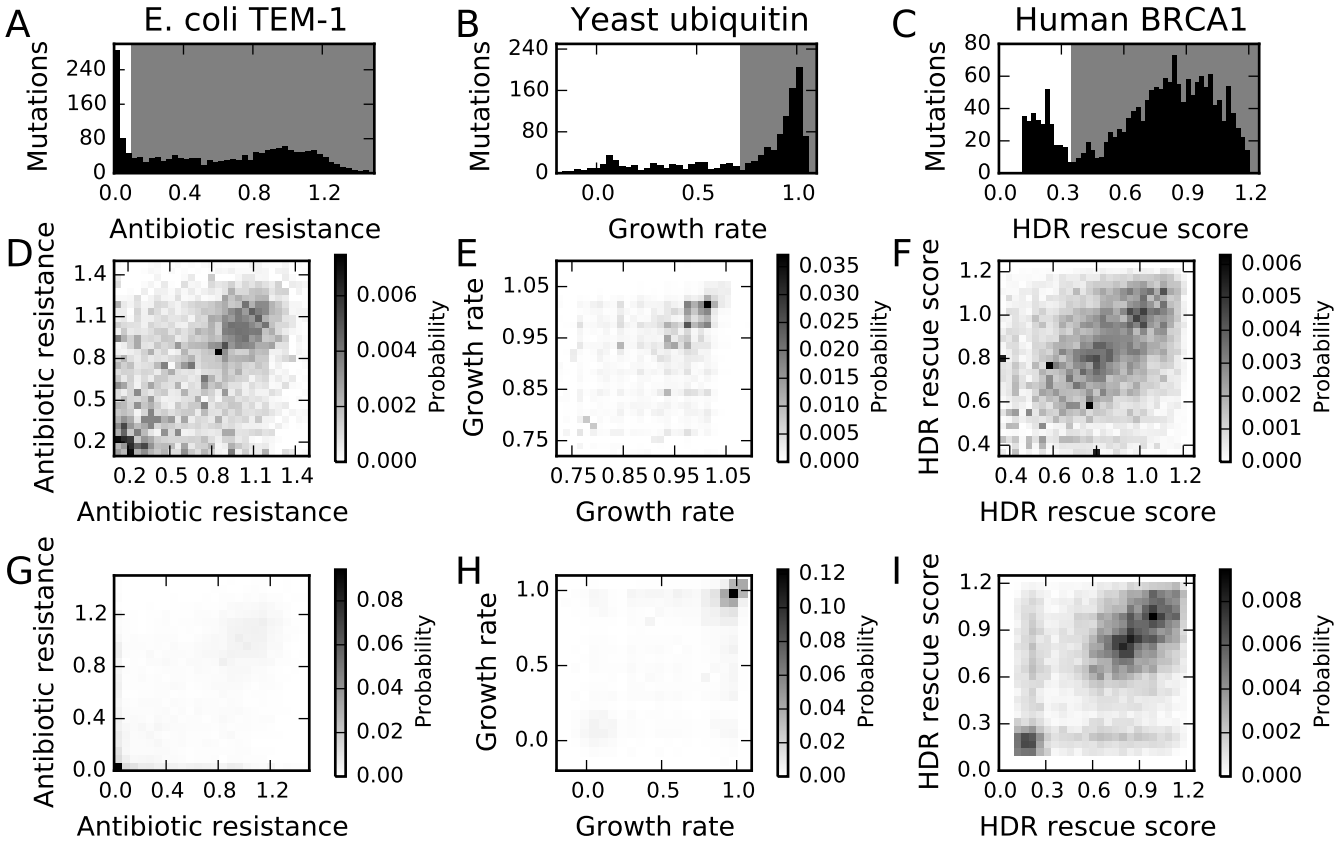

Figure S3: **Mutational scanning experiments.** A-C: Fitness proxy distributions for all three data sets. Shaded region indicates mutations considered in our primary analysis that focused on mutations with moderate effect (Table 1). D-F: Binned probability distribution of joint fitnesses for pairs of nonsynonymous mutations at the same site, restricted to fitnesses considered in our primary analysis. G-I: As in D-F, but including all mutations.

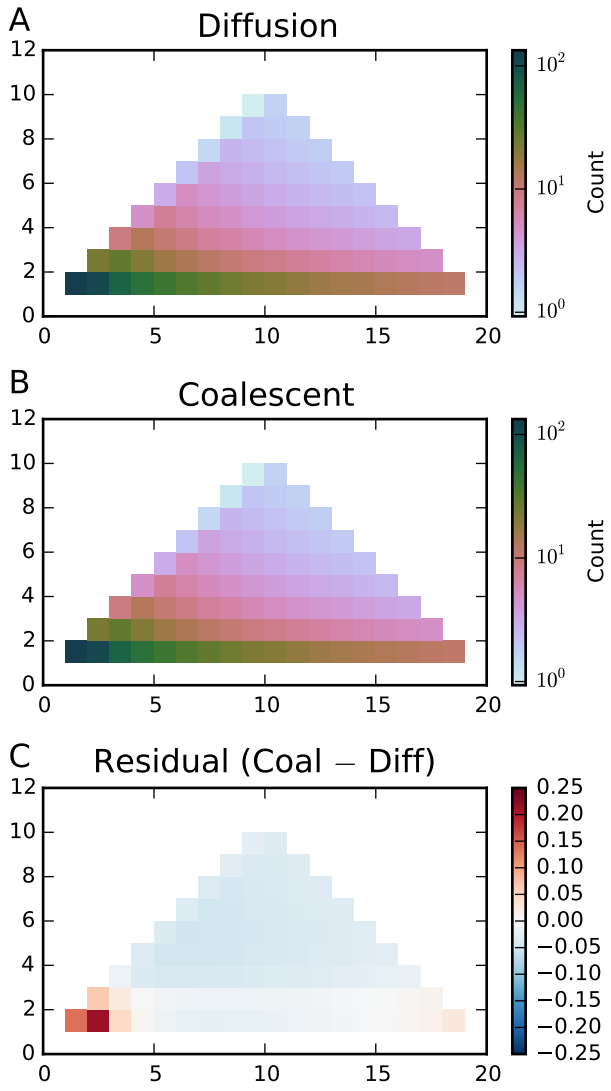

Figure S4: **Validation of numerics for neutral demography.** For a two-epoch demography in which the population doubles in size at a time $0.1 \times 2N_a$ generations before present, we compared our diffusion results with the coalescent method of Jenkins, Mueller, and Song (Jenkins *et al.*, 2014). A: Numerical diffusion solution with grid points [60, 80, and 100] and time steps [0.01, 0.001, and 0.0001]. B: Coalescent solution from 10,000 simulated trees (from Hudson's MS (Hudson, 2002)) used to calculate the moments of intercoalescent times. C. The residuals show good agreement between our numerical diffusion approach and the coalescent approach from Jenkins *et al.* (2014), with Poisson residuals calculated as (Coalescent − Diffusion)/$\sqrt{\text{Coalescent}}$, with the exception that our diffusion method slightly underestimates low-frequency alleles when compared with the coalescent approach.

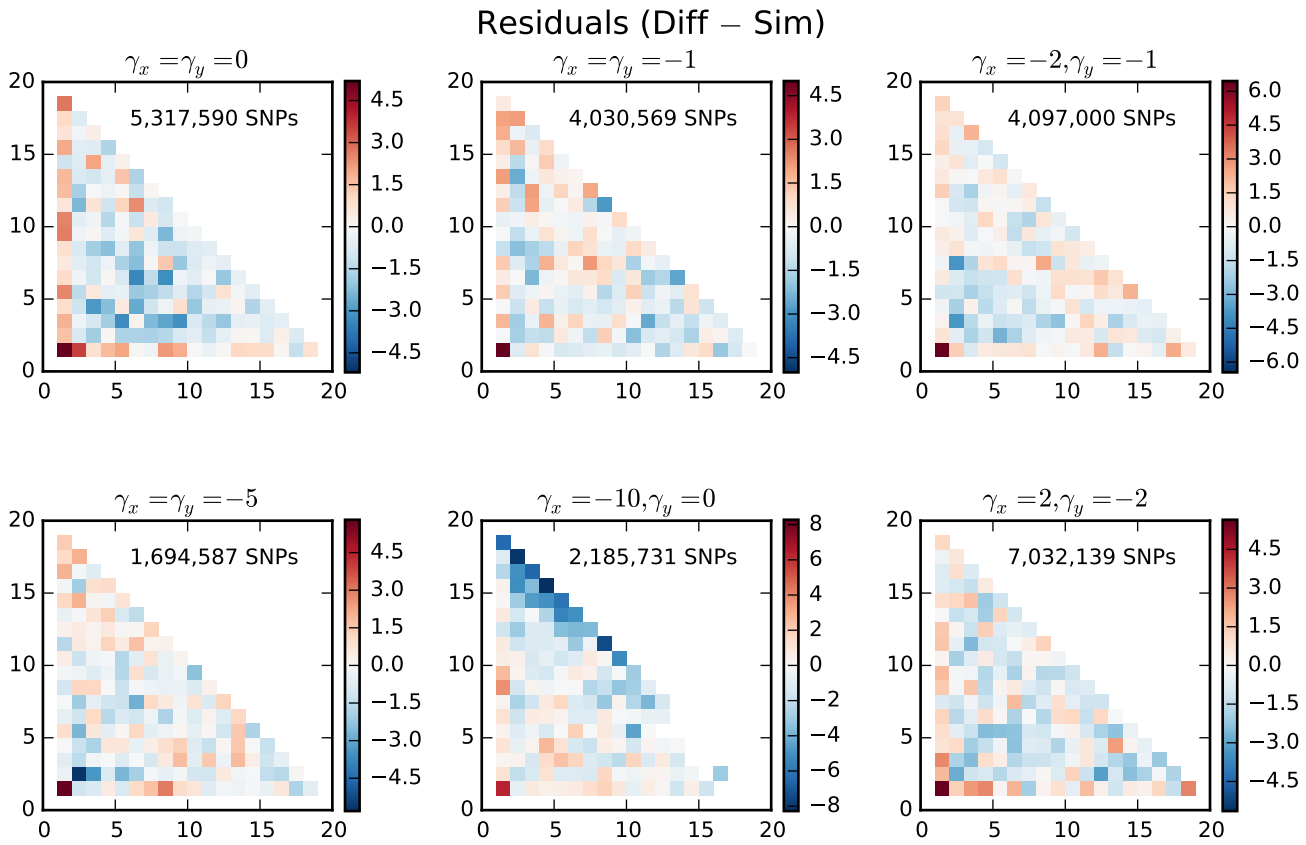

Figure S5: **Validation of numerics including selection.** Residuals between diffusion and discrete simulations at equilibrium for varying selection strengths. The diffusion-based spectra were scaled to match the magnitude of the simulation spectra. Poisson residuals were calculated as (Diffusion − Simulation)/$\sqrt{\text{Diffusion}}$. The number of SNPs in each comparison is labeled in each panel.

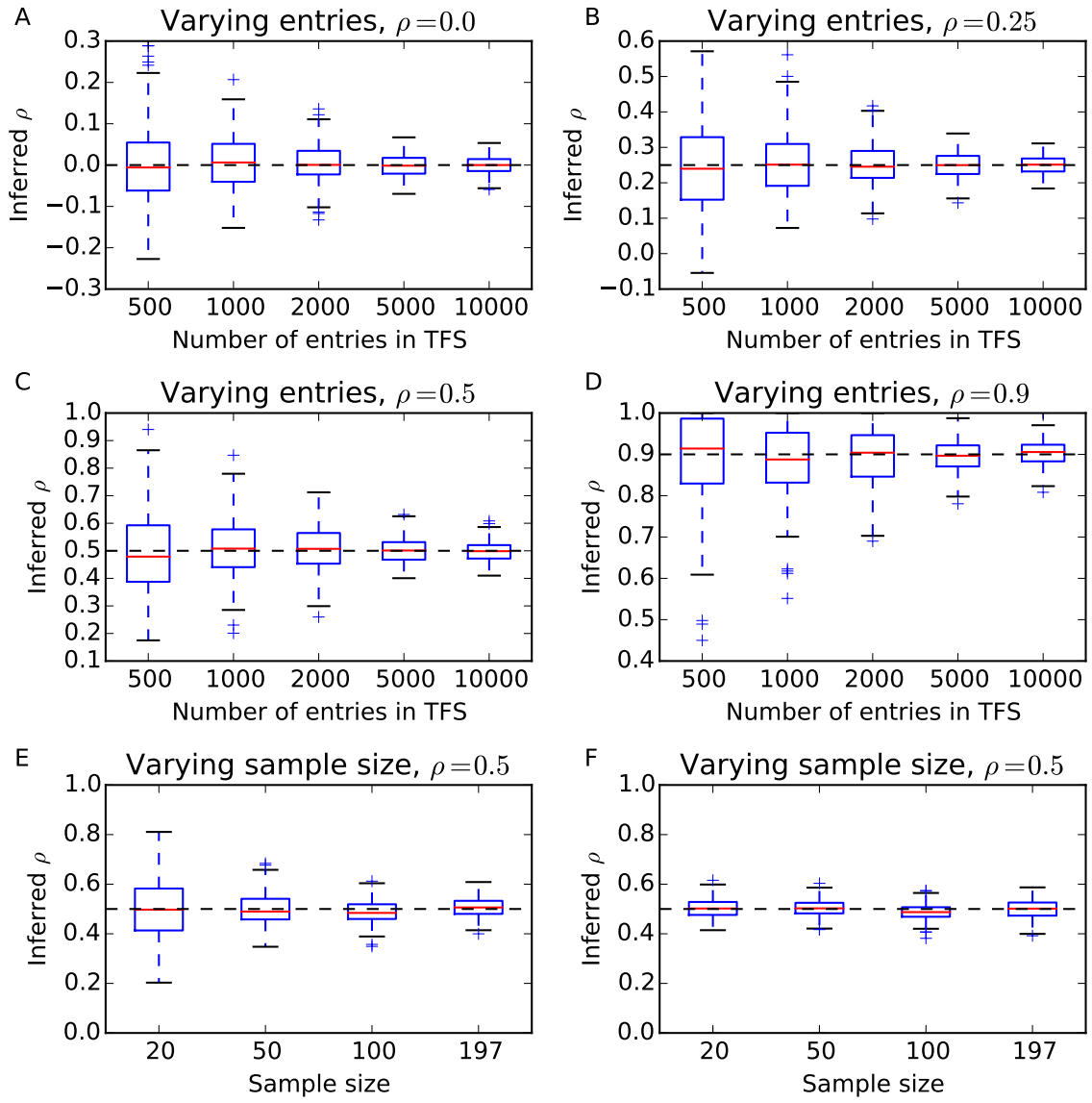

Figure S6: **Statistical power of selection correlation coefficient inference.** Plotted are the distributions of selection correlation coefficients inferred from simulated data. Red lines denote medians, boxes denote first and third quartiles. Data were simulated using our best-fit demography and marginal DFE from the *Drosophila* data, but with varying correlation coefficient ρ. For each configuration of ρ and data size, we simulated and fit 200 data sets. A-D: For a given ρ, we varied the number of observed triallelic loci among 197 sampled chromosomes. As the number of observed loci increases, the power to accurately infer ρ increases smoothly. E: For $\rho = 0.5$, we varied the number of sampled individuals, while holding θ constant so that 10,000 triallelic loci would be observed with 197 samples, and fewer would be observed with smaller samples. F: For $\rho = 0.5$, we varied the number of sampled individuals, while holding the number of observed triallelic loci constant at 10,000. In this case, power depends very little on sample size.

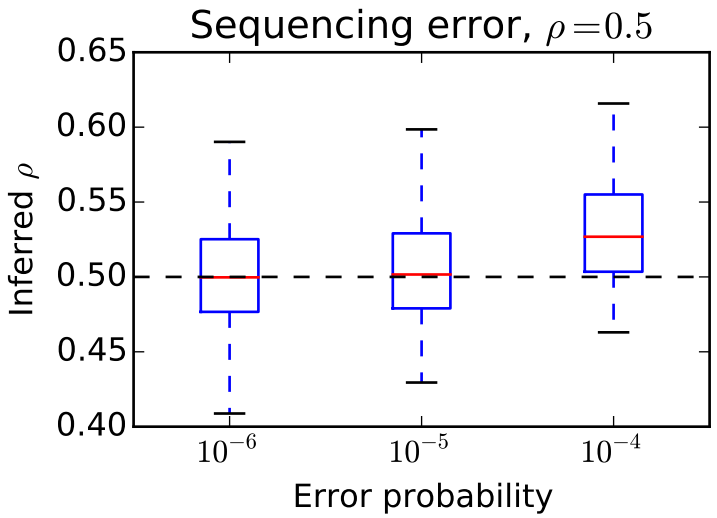

Figure S7: **Effect of sequencing error on selection correlation coefficient inference.** We simulated data under our best-fit demographic and marginal DFE model, with $\rho = 0.5$, including a model for sequencing error in which each individual genotype has probability ϵ of being miscalled (Johnson and Slatkin, 2008). Inference was unbiased for error rates $\epsilon \leq 10^{-5}$ and biased upward for larger error rates.

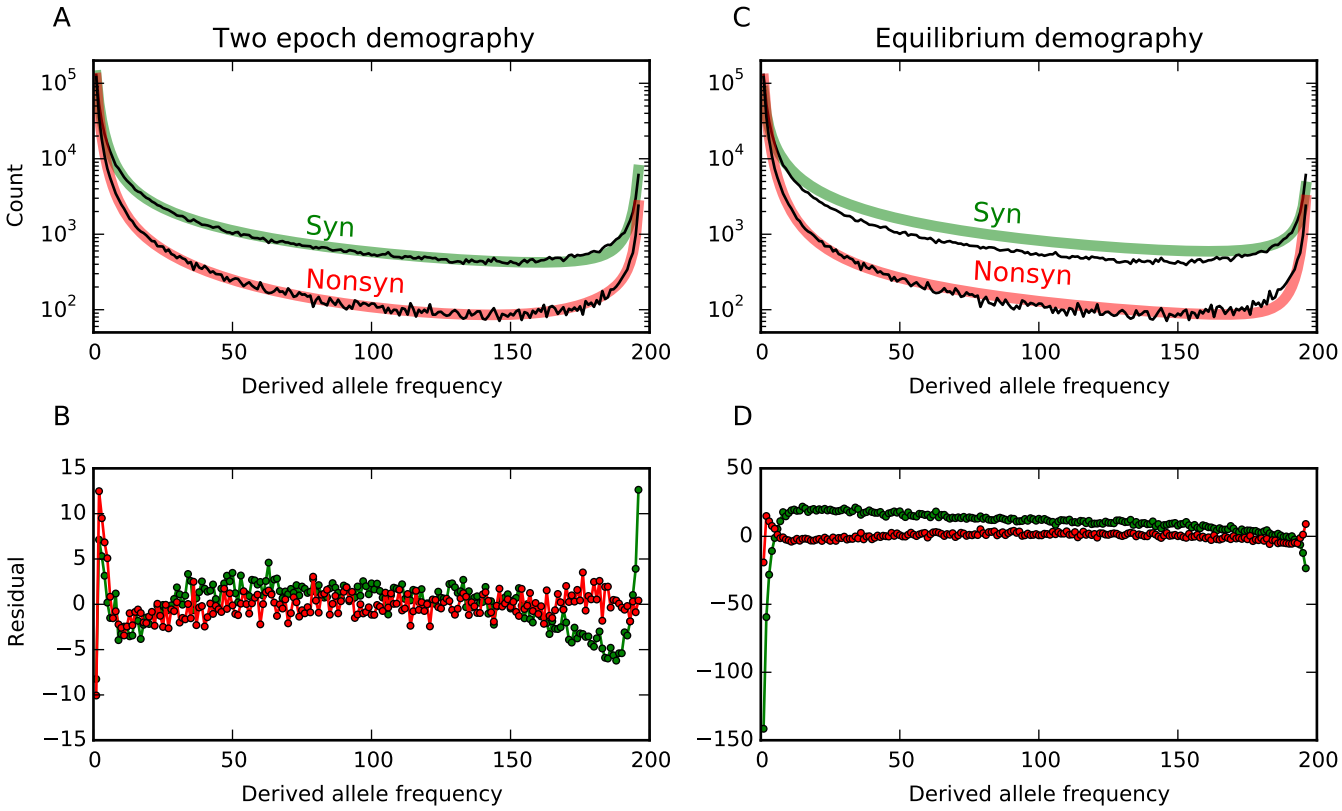

Figure S8: **Biallelic data fits for simpler demographic models.** Top panels show the data (thin black lines) and maximum-likelihood models (thick colored lines), and bottom panels show the Anscombe residuals between model and data. A&B: The two-epoch model fits the data more poorly than the three-epoch model. Our resulting inferences of the marginal DFE under this model are similar to under the three-epoch model, but the inferred value of ρ is notably lower. C&D: The equilibrium demographic model does not fit the biallelic data well, and both the marginal and bivariate DFE inferences differ substantially from the three-epoch model.

Table S1: **Maximum-likelihood estimates from *Drosophila* data.** Point estimates and confidence intervals for our three epoch demographic model (Fig. 2B), the univariate distribution of fitness effects for nonsynonymous mutations (Fig. 2C), and the bivariate DFE from triallelic data (Fig. 3).

		Estimates		
Model	Parameter	Max Likelihood	95% CI	Log-likelihood
3-epoch	ν_1	1.63	1.47−1.79	-1334
	ν_2	3.15	3.02−3.28	
	τ_1	0.12	0.10−0.14	
	τ_2	0.044	0.035−0.054	
	misid	0.052	0.051−0.053	
DFE	μ	5.42	5.23−5.62	-853.3
	σ	3.36	3.25−3.47	
	γ_+	39.9	35.7−44.1	
	p_+	0.0079	0.0069−0.0088	
	misid	0.014	0.013−0.015	
Correlation	ρ	0.51	0.44 − 0.59	-4433
	misid	0.029	0.023 − 0.34	

Table S2: **Estimated proportion of mutations in $N_e s$ ranges.** Eyre-Walker and Keightley (2009) used a gamma distribution for the DFE, and Kousathanas and Keightley (2013) used a lognormal DFE, both only for negative selection. The two analyses in Schneider et al. (2011) considered the strength and proportion of positively selected mutations.

	proportion in $N_e s$ range					
	≤ -100	$(-100, -10]$	$(-10, -1]$	$(-1, 0]$	> 0	$N_e s_+$
This study (lognormal, with positive selection)	0.51	0.25	0.15	0.079	0.0079	20
Eyre-Walker and Keightley (2009)	0.70	0.17	0.07	0.06	–	–
Kousathanas and Keightley (2013)	0.78	0.11	0.064	0.044	–	–
Schneider *et al.* (2011) (Shapiro *et al.* (2007) dataset)	–	–	–	–	0.0096	4.5
Schneider *et al.* (2011) (Callahan *et al.* (2011) dataset)	–	–	–	–	0.018	5.7

Table S3: **Maximum-likelihood parameters from simpler demographic models.** The two-epoch demographic model fit the synonymous data somewhat worse than the three-epoch model, and the equilibrium demography fit the data very poorly (Fig. S8). The fits to the nonsynonymous biallelic and triallelic data were also worse for the two simpler demographic models (compare log-likelihoods with Table S1). The two epoch demography fits the data more poorly for low and high frequency alleles than the three epoch demography (Fig. S8, and the DFE is skewed toward neutrality for negatively selected sites and the strength of positive selection is higher than the three epoch DFE. This causes more high frequency alleles to be present in the triallelic frequency spectrum, which tends to push the inferred ρ to lower values (Fig. 1).

		Estimates		
Model	Parameter	Max Likelihood	95% CI	Log-likelihood
2-epoch, demography	ν	2.73	$2.70 - 2.75$	-1499
	τ	0.091	$0.089 - 0.093$	
	misid	0.051	$0.050 - 0.052$	
2-epoch, DFE	μ	5.28	$5.25 - 5.30$	-979.3
	σ	3.03	$2.98 - 3.07$	
	γ_+	72.2	$69.0 - 75.4$	
	p_+	0.0091	$0.0084 - 0.0097$	
	misid	0.011	$0.010 - 0.012$	
2-epoch, correlation	ρ	0.31	$0.27 - 0.35$	-4451
	misid	0.025	$0.023 - 0.027$	
Equilibrium, DFE	μ	4.40	$4.39 - 4.42$	-1619
	σ	1.19	$1.17 - 1.21$	
	γ_+	0.0		
	p_+	0.062	$0.061 - 0.063$	
	misid	0.024	$0.023 - 0.024$	
Equilibrium, correlation	ρ	0.20	$0.17 - 0.22$	-4480
	misid	0.056	$0.050 - 0.062$	

Table S4: **Maximum likelihood estimates for a gamma-distributed DFE.** The univariate and bivariate gamma distributions fit the biallelic nonsynonymous *Drosophila* data worse than the lognormal distribution (Table S1). The correlation coefficient in the gamma distribution is in terms of the selection coefficient, but in the lognormal distribution it is in terms of the logarithm of the selection coefficient, so it is difficult to directly compare the inferred coefficients.

		Estimates		
Model	Parameter	Max Likelihood	95% CI	Log-likelihood
Univariate gamma	α	0.384	$0.379 - 0.389$	-871
	β	1401	$1339 - 1463$	
	γ_+	97.0	$82.3 - 111.7$	
	p_+	0.00426	$0.00396 - 0.00456$	
	misid	0.0133	$0.0122 - 0.0144$	
Gamma correlation	ρ	0.975	$0.966 - 0.983$	-4464
	misid	0.0197	$0.0146 - 0.0248$	

Table S5: **Maximum likelihood estimates for the refit DFEs for subsets of the data based on solvent accessibility and structural disorder.** We refit the correlation coefficient ρ to subsets of the triallelic fruit fly data. We first refit the univariate DFE to the biallelic nonsynonymous data, and we fixed $\gamma_+ = 39.9$ for all fits. In the biallelic fits, we inferred stronger negative selection for buried and ordered amino acids, as expected and described in the main text.

		Estimates	
Subset	Parameter	Max Likelihood	95% CI range
Exposed	μ	4.26	4.22−4.31
	σ	2.93	2.85−3.01
	p_+	0.0132	0.0118−0.0146
Buried	μ	7.06	6.98−7.14
	σ	3.53	3.45−3.60
	p_+	0.00352	0.00308−0.00396
Disordered	μ	4.14	4.11−4.17
	σ	3.00	2.95−3.05
	p_+	0.0111	0.0102−0.0120
Ordered	μ	6.42	6.37−6.47
	σ	3.43	3.37−3.48
	p_+	0.00589	0.00547−0.00631

File S1.

Triallelic solution numerics

Our numerical solution to Eq. 5 used finite differences, sequentially applying an alternating direction implicit (ADI) method for the non-mixed derivative terms and an explicit method for the mixed term. We used a uniform grid in x and y with grid spacing $\Delta = 1/(N-1)$, where N is the number of grid points in each direction. For the non-mixed terms, we discretized Eq. 5 as

$$\begin{aligned}\frac{\phi_{i,j}^{n+1} - \phi_{i,j}^{n}}{\Delta\tau} =& \frac{1}{2\nu}\frac{1}{\Delta}\left(\frac{V_{i+1}\phi_{i+1,j}^{n+1} - V_i\phi_{i,j}^{n+1}}{\Delta} - \frac{V_i\phi_{i,j}^{n+1} - V_{i-1}\phi_{i-1,j}^{n+1}}{\Delta}\right) \\ &- \frac{1}{2}\frac{1}{\Delta}\left(M_{i+1/2,j}\frac{\phi_{i,j}^{n+1} + \phi_{i+1,j}^{n+1}}{2} - M_{i-1/2,j}\frac{\phi_{i-1,j}^{n+1} + \phi_{i,j}^{n+1}}{2}\right)\end{aligned} \tag{S1}$$

where $V_i = x_i(1-x_i)$ and $M_{i,j} = \tilde{\gamma}_{i,j}x_i(1-x_i)$. The scheme for the y direction was equivalent. For the mixed term, we used the explicit discretization

$$\frac{\phi_{i,j}^{n+1} - \phi_{i,j}^{n}}{\Delta\tau} = -\frac{(C\phi^n)_{i+1,j+1} - (C\phi^n)_{i+1,j-1} - (C\phi^n)_{i-1,j+1} + (C\phi^n)_{i-1,j-1}}{4\Delta^2} \tag{S2}$$

where $C_{i,j} = x_i y_j$. The implicit (Eq. S1) and explicit (Eq. S2) schemes were sequentially applied each time step, along with the injection of density to account for new triallelic loci arising in the population due to mutation.

Because it does not capture the x, y covariation, the ADI method pushes excess density to the absorbing diagonal boundary (Fig. S1). To avoid this excess loss of density, we integrated ϕ using the ADI (Eq. S1) and explicit mixed derivative (Eq. S2) methods only for those grid points up to those immediately adjacent to the diagonal boundary but not including the grid points lying on the boundary. For points near the boundary, we calculated the amount of density that should be lost to the diagonal boundary from each grid point and explicitly removed that density each time step. The amount of density to be removed was numerically calculated by integrating the one dimensional diffusion equation forward in time by $\Delta\tau$ with initial point mass at frequency $x + y$, and then checking the relative amount of density that fixed with frequency one.

Simulations were initialized from the equilibrium neutral solution

$$\phi(x, y) = \frac{\theta}{x\,y}. \tag{S3}$$

For $\gamma \neq 0$, the system was integrated for $\tau = 20$ time units to reach equilibrium before applying demographic events.

To model the influx of density due to mutation, we coupled our numerical triallelic solution to models for the biallelic population frequency spectrum ϕ^{bx} and ϕ^{by}. We calculated these ϕ^{b} using ∂a∂i (Gutenkunst *et al.*, 2009). Each timestep we then added density to the first row of grid points on the interior of the domain near the horizontal or vertical axes. For example, along the y axis we added to $\phi_{1,j}$ the amount

$$\frac{1}{\Delta^2}\Delta\tau\frac{\theta}{2}\phi_j^{\mathrm{by}}. \tag{S4}$$

The diffusion equation applies for $N_e \to \infty$, so we extrapolate on the grid size to approximate the solution for $\Delta \to 0$. For numerical accuracy, we required more grid points than the number of samples in our study. There were 197 sequenced *Drosophila* individuals, so for our application we used $N = 250, 270$, and 290 grid points and quadratic extrapolation. We also found an increase in accuracy by extrapolating on the timestep $\Delta\tau$, and we used quadratic extrapolation on $\Delta\tau = 0.005, 0.001$, and 0.0002.

These methods were implemented in Python, using the NumPy (Oliphant, 2006), Scipy (Jones *et al.*, 2001–), and Matplotlib (Hunter, 2007) libraries, and we used Cython (Behnel *et al.*, 2011) to write some routines as a C-extension to Python for speed. The simulation code is packaged with ∂a∂i and available at `http://bitbucket.org/RyanGutenkunst/dadi`.

Validation of triallelic simulations

To validate our triallelic diffusion simulations, we first compared to the neutral coalescent results of Jenkins, Mueller, and Song (Jenkins *et al.*, 2014) for a number of simple two- and three-epoch demographies at neutrality, i.e. $\gamma_x = \gamma_y = 0$. We calculated the entries for the coalescent-based spectrum using Eq. 8 in Jenkins *et al.* (2014). Our diffusion simulations agreed well with these coalescent results (Fig. S4). Both the diffusion calculation and coalescent calculations took less than 30 seconds, although generating the trees needed for the coalescent calculation took tens of minutes.

To validate our triallelic diffusion simulations with selection, we compared to custom Wright-Fisher simulations implemented in Python (Figure S5). Each generation, new triallelic loci entered the population at constant rate, with frequency $(1/2N, y)$ or $(x, 1/2N)$, where frequencies x or y were sampled from the background equilibrium biallelic frequency spectrum corresponding to the selection coefficients of the two alleles. The expected number of new triallelic loci per generation depended on the selection coefficient of the background allele, as the magnitude of the biallelic density function varies with γ. We approximated these population-level density functions using $2N$ grid points. Each generation, the allele frequencies of each locus were updated using multinomial sampling, with the expected frequencies $x_{\text{next}} = x(1 + s_x)/(1 + xs_x + ys_y)$ and $y_{\text{next}} = y(1+s_y)/(1+xs_x+ys_y)$. After each generation, we dropped loci for which one of the alleles was lost, rendering the site no longer triallelic. We used population sizes of $N = 1000$ and ran each simulation for $20N$ generations to reach equilibrium. We then populated the frequency spectrum by sampling 20 copies of each locus every 100 generations, for $20N$ generations. We ran 200 independent instances of the simulation for each set of selection coefficients and combined the frequency spectra in order to increase the number of independently sampled triallelic loci. Our diffusion simulations agreed well with these discrete simulations (Fig. S5).

Bivariate lognormal distribution

For biallelic data, we used the univariate lognormal distribution to model the negatively selected portion of the nonsynonymous DFE, along with a point mass for some proportion of positive selection. The lognormal distribution takes the form

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} \tag{S5}$$

and we assume that with probability p_+, a novel nonsynonymous mutation will have selection coefficient $\gamma_+ > 0$, and with probability $1 - p_+$, it will have selection coefficient pulled from the density function $f(x)$. The bivariate extension of the lognormal distribution has density function

$$g(x, y) = \frac{1}{2\pi xy\sigma_x\sigma_y\sqrt{1-\rho}} e^{-Q/2} \tag{S6}$$

with

$$Q = \frac{1}{1-\rho^2}\left[\left(\frac{\ln x - \mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{\ln x - \mu_x}{\sigma_x}\right)\left(\frac{\ln y - \mu_y}{\sigma_y}\right) + \left(\frac{\ln y - \mu_y}{\sigma_y}\right)^2\right] \tag{S7}$$

where μ_x, μ_y are the marginal means, and σ_x, σ_y are the marginal standard deviations, and ρ is the correlation coefficient. In our application $\mu_x = \mu_y = \mu$, and $\sigma_x = \sigma_y = \sigma$.

We now describe how the DFE with a point mass of positive selection extends to the bivariate DFE. For a triallelic locus, assume that the first derived allele to arise has selection coefficient γ_x. This selection coefficient is γ_+ with probability p_+, or has $\gamma_x < 0$ sampled from the univariate lognormal distribution given by $f(x)$ with probability $1 - p_+$. The selection coefficient γ_y of the second derived allele at this locus is conditional on γ_x. The relative weights for the regimes of the signs of the two selection coefficients (with $\gamma_x, \gamma_y < 0$; $\gamma_x, \gamma_y > 0$; or one selection coefficient positive and the other negative) depends on both p_+ and ρ. If $\gamma_x = \gamma_+$, then $\gamma_y = \gamma_+$ with probability $\rho(1 - p_+) + p_+$, or γ_y is pulled from the univariate lognormal distribution with mean μ and standard deviation σ with probability $(1 - p_+)(1 - \rho)$. If ρ is larger, the

probability that $\gamma_y = \gamma_+$ as well increases, as expected. On the other hand, if $\gamma_x < 0$, having been pulled from the lognormal distribution with parameters μ and σ, then $\gamma_y = \gamma_+$ with probability $p_+(1-\rho)$, and $\gamma_y < 0$ with probability $(1-p_+) + p_+\rho$, pulled from the univariate lognormal distribution with $\tilde{\mu}$ and $\tilde{\sigma}$. The new parameters $\tilde{\mu}$ and $\tilde{\sigma}$ are found by holding γ_x constant in the bivariate lognormal distribution and rearranging to find the univariate distribution for γ_y:

$$\begin{aligned}\tilde{\sigma} &= \sigma\sqrt{1-\rho},\\ \tilde{\mu} &= \mu + \rho\tilde{\sigma}B,\\ B &= \frac{\ln(-\gamma_x) - \mu}{\sigma\sqrt{1-\rho^2}}.\end{aligned}$$

Bivariate gamma distribution

We also performed our analysis using a gamma distributed DFE instead of a lognormal DFE. The bivariate gamma distribution we used was introduced by Kibble (1941). The gamma distribution takes the form

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \tag{S8}$$

where α and β are the shape and scale parameters, respectively. Consider two marginal gamma distributions f_1 with shape and scale parameters α_1 and β_1, and f_2 with α_2, β_2, and let $\alpha = \sqrt{\alpha_1\alpha_2}$. For correlation coefficient ρ, the bivariate gamma distribution that we used then takes the form

$$g(x,y) = \frac{(\beta_1\beta_2)^\alpha}{(1-\rho)\Gamma(\alpha)} \left(\frac{xy}{\rho\beta_1\beta_2}\right)^{\frac{\alpha-1}{2}} \exp\left(-\frac{\beta_1 x + \beta_2 y}{1-\rho}\right) I_{\alpha-1}\left(\frac{2\sqrt{\rho\beta_1\beta_2 xy}}{1-\rho}\right) \tag{S9}$$

where I_α is the modified Bessel function of the first kind of order α, given by

$$I_\alpha(z) = \sum_{k=0}^{\infty} \frac{(z/2)^{2k+\alpha}}{\Gamma(k+\alpha+1)k!}. \tag{S10}$$

Subsets of trialleic variation

When considering similar or dissimilar derived amino acids, we refit the correlation coefficient using the same univariate DFE as the full data. For the solvent accessibility and disorder analysis, we refit the univariate DFE, although for consistency with the whole-genome analysis we fixed $\gamma_+ = 39.9$.

To predict residue solvent accessibility and structural disorder using SPINE-D (Zhang *et al.*, 2012), we first matched RefSeq IDs from Annovar (Wang *et al.*, 2010) to Ensembl release 81 (Cunningham *et al.*, 2014) gene and transcript IDs. For the 95% of proteins for which the Ensembl CDS sequence perfectly matched the SNP states reported by Annovar, we then ran SPINE-D (Zhang *et al.*, 2012) on the longest Ensembl transcript.

Mutational scanning data

Firnberg *et al.* (2014) studied 286 positions in *E. coli* TEM-1 β-Lacatamase. Their proxy for fitness was resistance to ampicillin, relative to the ancestral sequence. They reported fitnesses at the codon level, so we restricted our analysis to mutations that were a single nucleotide change from the ancestral sequence. The distribution of measured fitnesses was strongly skewed toward zero (Fig. S3A). To exclude severely deleterious mutations, in our primary analysis (Table 1) we considered only mutations with fitness > 0.1 (Fig. S3A,D). If we included all mutations (Fig. S3G), we inferred a fitness correlation coefficient of $0.42-0.58$.

Roscoe *et al.* (2013) studied 75 positions in *S. cerevisiae* ubiquitin, assaying fitness as growth rate relative to the ancestral sequence. They reported data at the amino acid level, so we considered mutations to all possible amino acids. To focus on the bulk of the fitness distribution, we restricted our primary analysis to only mutations with fitness > 0.72 (Fig. S3B,E). If we included all mutations (Fig. S3H), we inferred a fitness correlation coefficient of $0.34-0.68$.

Starita *et al.* (2015) studied 303 positions in the RING domain of human BRCA1 protein. Their proxy for fitness was a homology-directed DNA repair (HDR) score, based on combined measurements of ubiquitination activity assayed by phage display and binding to the RING domain of BARD1 assayed by yeast two-hybrid. To focus on the bulk of the fitness distribution, we restricted our primary analysis to consider only mutations with fitness > 0.35 (Fig. S3C,F). If we included all mutations (Fig. S3I), we inferred a fitness correlation coefficient of $0.29-0.59$.

References

Behnel, S., R. Bradshaw, C. Citro, L. Dalcin, D. Seljebotn, and K. Smith, 2011 Cython: the best of both worlds. Comput Sci Eng **13**: 31 –39.

Callahan, B., R. a. Neher, D. Bachtrog, P. Andolfatto, and B. I. Shraiman, 2011 Correlated evolution of nearby residues in Drosophilid proteins. PLoS Genet **7**: e1001315.

Cunningham, F., M. R. Amode, D. Barrell, K. Beal, K. Billis, *et al.*, 2014 Ensembl 2015. Nucleic Acids Res **43**: D662–D669.

Eyre-Walker, A. and P. D. Keightley, 2009 Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol **26**: 2097–2108.

Firnberg, E., J. W. Labonte, J. J. Gray, and M. Ostermeier, 2014 A comprehensive, high-resolution map of a gene's fitness landscape. Mol Biol Evol **31**: 1581–1592.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet **5**: e1000695.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**: 337–338.

Hunter, J. D., 2007 Matplotlib: A 2d graphics environment. Comput Sci Eng **9**: 90–95.

Jenkins, P. A., J. W. Mueller, and Y. S. Song, 2014 General triallelic frequency spectrum under demographic models with variable population size. Genetics **196**: 295–311.

Johnson, P. L. F. and M. Slatkin, 2008 Accounting for bias from sequencing error in population genetic estimates. Mol Biol Evol **25**: 199–206.

Jones, E., T. Oliphant, P. Peterson, *et al.*, 2001– SciPy: Open source scientific tools for Python.

Kibble, W. F., 1941 A two-variate gamma type distribution. Sankhya **5**: 137–150.

Kousathanas, A. and P. D. Keightley, 2013 A comparison of models to infer the distribution of fitness effects of new mutations. Genetics **193**: 1197–1208.

Oliphant, T. E., 2006 *Guide to NumPy*. Provo, UT.

Roscoe, B. P., K. M. Thayer, K. B. Zeldovich, D. Fushman, and D. N. A. Bolon, 2013 Analyses of the effects of all ubiquitin point mutants on yeast growth rate. J Molec Biol **425**: 1363–1377.

Schneider, A., B. Charlesworth, A. Eyre-Walker, and P. D. Keightley, 2011 A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics **189**: 1427–1437.

Shapiro, J. A., W. Huang, C. Zhang, M. J. Hubisz, J. Lu, *et al.*, 2007 Adaptive genic evolution in the Drosophila genomes. Proc Natl Acad Sci U S A **104**: 2271–2276.

Starita, L. M., D. L. Young, M. Islam, J. O. Kitzman, J. Gullingsrud, *et al.*, 2015 Massively parallel functional analysis of BRCA1 RING domain variants. Genetics **200**: 413–422.

Wang, K., M. Li, and H. Hakonarson, 2010 ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res **38**: 1–7.

Zhang, T., E. Faraggi, B. Xue, A. K. Dunker, V. N. Uversky, and Y. Zhou, 2012 SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. J Biomolec Struct Dynam **29**: 799–813.