Article:

Barton, H.J. and Zeng, K. (2018) New methods for inferring the distribution of fitness effects for INDELs and SNPs. Molecular Biology and Evolution, 35 (6). pp. 1536-1546. ISSN 0737-4038

https://doi.org/10.1093/molbev/msy054

New methods for inferring the distribution of fitness effects for INDELs and SNPs

Henry J. Barton[1] and Kai Zeng[*,1]

[1] Department of Animal and Plant Sciences, University of Sheffield, Sheffield, S10 2TN, United Kingdom

***Corresponding author:** E-mail: k.zeng@sheffield.ac.uk

Abstract

Small insertions and deletions (INDELs; ≤50bp) are the most common type of variability after SNPs. However, compared to SNPs, we know little about the distribution of fitness effects (DFE) of new INDEL mutations and how prevalent adaptive INDEL substitutions are. Studying INDELs has been difficult partly because identifying ancestral states at these sites is error-prone and misidentification can lead to severely biased estimates of the strength of selection. To solve these problems, we develop new maximum likelihood methods, which use polymorphism data to simultaneously estimate the DFE, the mutation rate, and the misidentification rate. These methods are applicable to both INDELs and SNPs. Simulations show that they can provide highly accurate results. We applied the methods to an INDEL polymorphism dataset in *Drosophila melanogaster*. We found that the DFE for polymorphic INDELs in protein-coding regions is bimodal, with the variants being either nearly neutral or strongly deleterious. Based on the DFE, we estimated that 71.5% – 83.7% of the INDEL substitutions that took place along the *D. melanogaster* lineage were fixed by positive selection, which is comparable to the prevalence of adaptive substitutions at non-synonymous sites. The new methods have been implemented in the software package `anavar`.

Key words: Distribution of fitness effects, insertions and deletions, single nucleotide polymorphism, polarisation error

Article

Introduction

New mutations can have a range of effects on an organism's fitness, ranging from being strongly harmful, through being only slightly deleterious, to being neutral, and finally on to being either mildly or highly beneficial. The relative frequencies of mutations with different selective effects is known as the distribution of fitness effects (DFE). The DFE is an important parameter as it is required for addressing many fundamental questions (Eyre-Walker and Keightley, 2007). Examples include understanding determinants of the efficacy of natural selection (Corcoran *et al.*, 2017; Galtier, 2016), the genetic basis of polygenic traits (Zuk *et al.*, 2014), and the evolutionary advantage of sex and recombination (Hartfield and Keightley, 2012).

Taking advantage of the massive increase in data availability, many methods have been proposed for estimating the DFE using polymorphism data (Eyre-Walker and Keightley, 2009; Eyre-Walker *et al.*, 2006; Keightley and Eyre-Walker, 2007; Kim *et al.*, 2017; Kousathanas and Keightley, 2013; Tataru *et al.*, 2017). Their development in turn allows more reliable inferences about other important quantities such as α, the proportion of adaptive substitutions (Eyre-Walker and Keightley, 2009). However, all these methods are concerned with estimating the DFE for single nucleotide polymorphisms (SNPs). Consequently, much less is known about the DFE and α for other types of genetic variation such as small insertions and deletions (INDELs; $\leq$ 50bp), despite the fact that INDELs are the second most common type of variants (e.g., Montgomery *et al.*, 2013), and hence represent an important source of raw materials for selection to act on.

A major difficulty in studying INDELs lies with ancestral state identification. This requires multi-species genome alignments. However, INDELs occur disproportionately in repetitive genomic regions (Ananda *et al.*, 2013; Montgomery *et al.*, 2013), where alignment algorithms perform poorly (Earl *et al.*, 2014). Furthermore, there is evidence that homoplasy is a significant issue outside repetitive regions, probably due to the existence of cryptic INDEL mutation hotspots (Kvikstad and Duret, 2014). Thus ancestral state identification can be expected to be particularly error prone for INDELs. It is well established that misidenfication of ancestral states can lead to severely biased estimates of the strength of selection using the site-frequency spectrum (SFS) (Hernandez *et al.*, 2007). For SNPs, this difficulty can be avoided by using the folded SFS (e.g., Eyre-Walker *et al.*, 2006; Keightley and Eyre-Walker, 2007). However, to determine whether a length variant is an insertion or a deletion, we have to know what the ancestral state is, meaning that the issue of polarisation error is inherent for INDELs. As a result, applying existing methods for estimating the DFE to INDEL data may be liable to biases.

Another challenge is that the SFSs for insertions and deletions may be affected by polarisation errors to different extents. This is because when the ancestral state of an insertion segregating at low frequency is misidentified, it will be incorrectly inferred as a deletion segregating at high frequency (and vice versa). There is direct experimental evidence that the deletion mutation rate is higher than the insertion mutation rate (Besenbacher *et al.*, 2015; Keightley *et al.*, 2009; Schrider *et al.*, 2013; Yang *et al.*, 2015). This mutational bias means that there are more deletions segregating in the population than insertions. The larger number of deletions may lead to the SFS for insertions being disproportionally affected by polarisation errors (Figure 1). This asymmetry can cause the insertion SFS to have a more pronounced, but artificial, uptick at the high-frequency end, which can be misinterpreted as stronger positive selection on insertions over deletions. As pointed out by Kvikstad and Duret (2014), this methodological issue can, at least in principle, compromises the results of previous studies, which suggest that insertions are more likely to be under positive selection than deletions to prevent the genome size from unconstrained contraction caused by the mutational bias towards deletions (Parsch, 2003). Similarly, it will make it difficult to test the possibility that insertions have a higher fixation probability because they are favoured by insertion-biased gene conversion (Leushkin and Bazykin, 2013).

Towards resolving the confounding efforts ancestral state misidentification have on the study of INDELs, we propose new maximum likelihood methods for inferring the DFE using polymorphism data. These methods are based on recent studies on SNPs which show that polymorphism data contains enough information for simultaneous estimation of the mutation rate, the DFE, and the polarisation error rate (Glémin *et al.*, 2015; Tataru *et al.*, 2017). Our methods are more general than the existing methods in the following aspects. First, they can handle both INDELs and SNPs. Second, insertions and deletions can have different polarisation error rates, mutation rates, and DFEs. Third, for both INDELs and SNPs, the new methods allow the mutation and polarisation error rates to vary across the genome. Incorporating these heterogeneities may be particularly important for INDELs (Kvikstad and Duret, 2014). We carried out extensive simulations to examine the performance of the new methods. As an example, we applied the methods to an INDEL polymorphism dataset in *Drosophila melanogaster* we obtained by re-analysing the raw short-read data published by the *Drosophila* Population

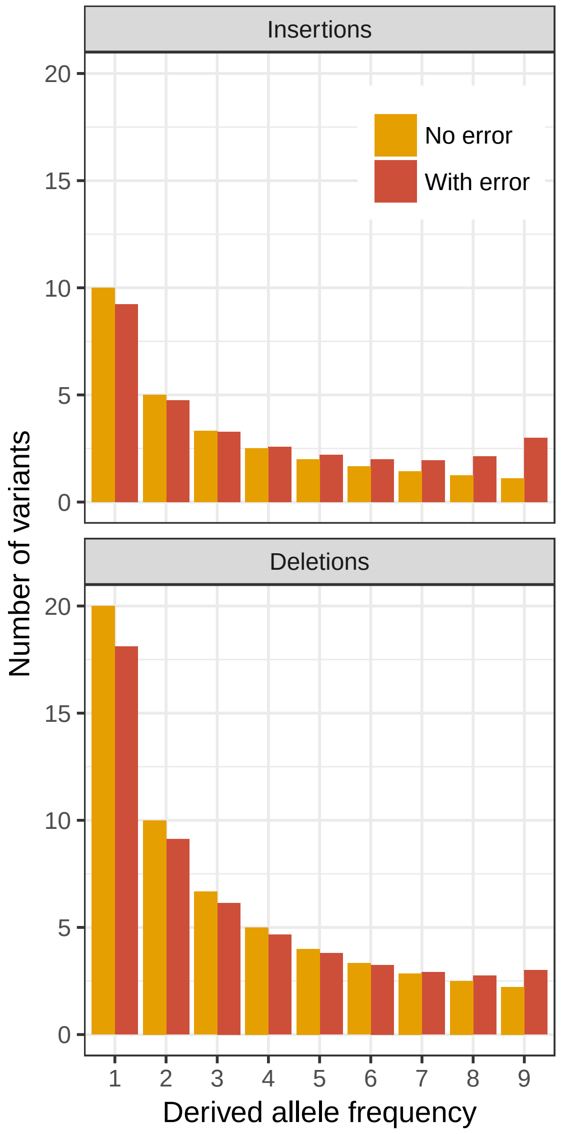

FIG. 1. The SFSs for insertions and deletions may be affected to different extents by polarisation errors. We assume that the population size is constant, that INDELs are neutral, and that the sample size is 10. In the genomic region under consideration, the total scaled mutation rate towards insertions, $4N_eum$, is 10, where N_e is the effective population size u is the insertion mutation rate per site per generation, and m is that size of the focal region. The total scaled mutation rate towards deletions is 20. The expected SFSs were generated using standard neutral theory. The SFSs with polarisation errors were generated by assuming that the ancestral state of an INDEL was wrongly identified with probability 0.1.

Genomics Project (Pool *et al.*, 2012). Through model comparisons, we tried to find the DFE that best described the observed pattern of INDEL polymorphism within protein-coding regions of the genome. Finally, using the best-fitting DFE, we estimated the proportion of INDEL substitutions fixed by positive selection (α).

New Approach

For ease of presentation, we will start with a description of the SNP models. The INDEL models will be presented later as an extension.

The SNP models

Consider a diploid population with effective size N_e. The size of the genomic region of interest is m base pairs, and the sample size is n.

The discrete model:

Assume that there are C different classes of sites in the focal region. These sites can be different with respect to their mutation rates, the fitness effects of new mutations, and polarisation error rates. This discrete model has several advantages. First, it does not assume that the DFE follows a specific probability distribution, and is therefore able to accommodate complex scenarios such as a multi-modal DFE (Kousathanas and Keightley, 2013). Second, by allowing the mutation and polarisation error rates to vary freely between site classes, the method can include situations whereby these two variables co-vary (e.g., hypermutable regions may have a higher polarisation error rate).

We assume that the mutation process can be approximated by the infinite-sites model. Let the total scaled mutation rate for sites of class c be $m\theta_c$, where $c \in \{1,2,...,C\}$ and $\theta_c = 4N_e u_c$. To understand u_c, consider an alternative formulation whereby the mutation rate for the c^{th} class of sites is v_c per site per generation, and sites of class c account for a fraction p_c of all sites in the focal region (i.e., $\sum_c p_c = 1$). We have $m\theta_c = mp_c 4N_e v_c$, which leads to $u_c = p_c v_c$. By using θ_c, we can perform searches for maximum likelihood estimates (MLEs) of the parameters without having to deal with the constraint $\sum_c p_c = 1$. Define

$$\theta = \sum_{c=1}^{C} \theta_c = 4N_e \sum_{c=1}^{C} p_c v_c. \tag{1}$$

Thus, θ is the average scaled mutation rate per site, and the total scaled mutation rate is $m\theta$. If the per-site mutation rate is uniform across the focal region (i.e., $v_i = v_j$ for $i \neq j$ and $1 \leq i,j \leq C$), then $\theta_c/\theta = p_c$.

To model selection, we assume that, for mutations arising at sites of class c, the fitnesses of the wild-type, heterozygote, and mutant homozygote genotypes are 1, 1 + s_c, and 1 + $2s_c$, respectively. The corresponding scaled selection coefficient γ_c is defined as $4N_e s_c$. Positive and negative γ_c values signify beneficial and deleterious mutations, respectively.

The site-frequency spectrum (SFS) for the c^{th} site class, which is defined as the expected number of polymorphic sites of size i (i.e., sites where the derived allele is represented i times; $1 \le i < n$), is given by

$$\Psi_{c,i} = m\theta_c \tau_i(\gamma_c) \tag{2}$$

where

$$\tau_i(\gamma) = \int_0^1 \binom{n}{i} x^i (1-x)^{n-i} \frac{1-e^{-\gamma(1-x)}}{x(1-x)(1-e^{-\gamma})} dx. \tag{3}$$

Polarisation errors distort the SFS. Specifically, when the ancestral state of a polymorphic site of size i is mis-identified, it will be regarded as a polymorphic site of size $n-i$. To model polarisation errors, we let ϵ_c be the probability that the ancestral state of a polymorphic site of class c is incorrectly identified (Glémin *et al.*, 2015). The final SFS for sites of class c is then

$$\psi_{c,i} = (1-\epsilon_c)\Psi_{c,i} + \epsilon_c \Psi_{c,n-i}. \tag{4}$$

In what follows, we refer to the SFS with and without the correction of polarisation errors as the corrected and uncorrected SFS, respectively. The corrected SFS for the focal region is simply the sum of all the contributions from the sites in different classes

$$\psi_i = \sum_{c=1}^{C} \psi_{c,i}. \tag{5}$$

Existing models either do not model polarisation error (Eyre-Walker and Keightley, 2009; Keightley and Eyre-Walker, 2007; Kim *et al.*, 2017) or assume that the error rate is constant across the focal region (Glémin *et al.*, 2015; Tataru *et al.*, 2017). The model described above is therefore more general. Allowing variation in the polarisation error rate can be important. For instance, sites under stronger selective constraints tend to evolve slower, and are less likely to be polarised incorrectly due to homoplasy. It should, however, be noted that, when $\gamma_c \equiv \gamma$ for $\forall c \in \{1,2,...,C\}$, not all the parameters are identifiable. To see this, we rewrite (5) as

$$\psi_i = m\sum_{c=1}^{C}(1-\epsilon_c)\theta_c\tau_i(\gamma) + m\sum_{c=1}^{C}\epsilon_c\theta_c\tau_{n-i}(\gamma). \tag{6}$$

Appealing to (1) and defining ϵ^* such that

$$\epsilon^*\theta = \sum_{c=1}^{C}\epsilon_c\theta_c \tag{7}$$

we can rewrite (6) as

$$\psi_i = (1-\epsilon^*)m\theta\tau_i(\gamma) + \epsilon^* m\theta\tau_{n-i}(\gamma). \tag{8}$$

Thus, when there is no difference in fitness effects between mutations arising at sites of different classes, we cannot detect variation in the scaled mutation rate and polarisation error rate because the model reduces to one that depends on θ, γ and ϵ^*. This result has important implications for data analysis by pointing out that a model with a small number of site classes may provide an adequate description of the data even when the underlying biological process features complex variation in the mutation rate across the genome.

The continuous model:

Instead of assuming that the focal region is composed of several classes of sites, we can assume that the fitness effects of new mutations follows a continuous distribution characterised by parameters Ω. Let θ be the scaled mutation rate per site, and ϵ be the polarisation error rate. The uncorrected SFS becomes

$$\Psi_i = m\theta \int \tau_i(\gamma) f(\gamma|\Omega) d\gamma \tag{9}$$

where $f(\gamma|\Omega)$ is the probability density function. The corrected SFS is analogous to (4) with c in the subscripts omitted.

Although the modelling framework allows the DFE to follow arbitrary probability distribution (including those mixture distributions considered by Galtier (2016)), here we only consider the reflected Γ distribution, i.e., $-\gamma \sim \Gamma(a,b)$, where $\gamma \leq 0$ and a and b are the shape and scale parameters, respectively.

Parameter estimation:

Let $X = (x_1, x_2, ..., x_{n-1})$ represent the observed SFS, where x_i is the number of polymorphic sites of size i in the sample. Let Θ denote all the parameters in the model (i.e., θ_c, γ_c, and ϵ_c for $c \in \{1,2,...,C\}$ for the discrete model and θ, Ω, and ϵ for the continuous model). To obtain MLEs of Θ, we use the Poisson random field model (Bustamante *et al.*, 2001; Sawyer and Hartl, 1992). Omitting constants that have no effects on the shape of the likelihood surface, the log likelihood function is defined as

$$L(\Theta|X) = \sum_{i=1}^{n-1} \big(-\psi_i + x_i \ln(\psi_i)\big). \tag{10}$$

Controlling for demography:

We have so far assumed that the population is panmictic and of constant size N_e. To control for demography, we employ the method of Eyre-Walker *et al.* (2006). Take the continuous model as an example. First, we define augmented SFSs as

$$\begin{cases} \Psi_i^* = r_i \Psi_i & \text{(11a)} \\ \psi_i^* = (1-\epsilon)\Psi_i^* + \epsilon \Psi_{n-i}^* & \text{(11b)} \end{cases}$$

Next, a set of neutral variants is added to the model, which introduces two additional parameters $\theta^{(0)}$ and $\epsilon^{(0)}$, which are the scaled mutation rate per site and the polarisation error rate, respectively, for the neutral sites. Let $\Theta^{(0)}$ denote these new parameters and $X^{(0)}$ denote the neutral SFS. The log likelihood of the observed data can be calculated as

$$L(\Theta, \Theta^{(0)}, R|X, X^{(0)}) = L(\Theta, R|X) + L(\Theta^{(0)}, R|X^{(0)}) \tag{12}$$

where $R = (r_2, r_3, ..., r_{n-1})$ and the two log likelihood functions on the right-hand side are calculated in the same way as (10) with ψ_i replaced by ψ_i^*.

The above method for controlling for demography has been used extensively (Eyre-Walker *et al.*, 2006; Galtier, 2016; Glémin *et al.*, 2015; Jackson *et al.*, 2017; Muyle *et al.*, 2011; Tataru *et al.*, 2017). These previous efforts have gathered clear theoretical and empirical evidence that the method is robust against a wide range of demographic processes, as well as the effects caused by selection at linked sites (e.g., background selection and/or selective sweeps). For instance, in a recent analysis of selection on codon usage bias in *Drosophila*, Jackson *et al.* (2017) showed that the estimates of γ produced by an estimation method that corrects for demography using the r parameters as set out above closely matched those produced by another estimation method that considers an explicit one-step change in population size (see Figure 4A in Jackson *et al.* (2017)).

It should be noted that (12) accommodates the possibility that the focal region and the neutral region have different mutation rates. This is more general than several previous models (Eyre-Walker and Keightley, 2009; Keightley and Eyre-Walker, 2007; Kim *et al.*, 2017; Tataru *et al.*, 2017). However, it may be challenging to distinguish this model from one in which the two regions have the same mutation rate, but a proportion of new mutations in the focal region are so strongly deleterious that they make negligible contributions to the observed SFS.

The INDEL models

The discrete model:

First consider insertions. Assume that there are C^{ins} different classes of sites. The total scaled mutation rate towards insertions for sites of class c is $m\theta_c^{ins}$, and the fitness effect and polarisation error rate are γ_c^{ins} and ϵ_c^{ins}, respectively ($1 \leq c \leq C^{ins}$). The uncorrected SFS for insertions of class c can be calculated using (2), and is denoted by $\Psi_{c,i}^{ins}$. For deletions, we can similarly assume that there are C^{del} different classes of sites. The associated parameters are θ_d^{del}, γ_d^{del}, and ϵ_d^{del}, and the uncorrected SFS is denoted by $\Psi_{d,i}^{del}$ ($1 \leq d \leq C^{del}$).

When the ancestral state of a derived insertion of size i is misidentified, it will be wrongly identified as a deletion of size $n-i$, and vice versa for deletions (note that size in this context refers to the frequency of the derived allele, not the number of base pairs inserted or deleted). Thus, the corrected SFSs for insertions and deletions are

$$\begin{cases} \psi_i^{ins} = \sum_{c=1}^{C^{ins}} (1-\epsilon_c^{ins}) \Psi_{c,i}^{ins} + \sum_{d=1}^{C^{del}} \epsilon_d^{del} \Psi_{d,n-i}^{del} & \text{(13a)} \\ \psi_i^{del} = \sum_{d=1}^{C^{del}} (1-\epsilon_d^{del}) \Psi_{d,i}^{del} + \sum_{c=1}^{C^{ins}} \epsilon_c^{ins} \Psi_{c,n-i}^{ins} & \text{(13b)} \end{cases}$$

The continuous model:

For insertions, define the per-site scaled mutation rate and the polarisation error rate as θ^{ins} and ϵ^{ins}, respectively. The DFE for insertions is determined by parameters Ω^{ins}. For deletions, we similarly define the following parameters: θ^{del}, Ω^{del} and ϵ^{del}. Finally, the corrected SFSs are

$$\begin{cases} \psi_i^{ins} = (1-\epsilon^{ins}) \Psi_i^{ins} + \epsilon^{del} \Psi_{n-i}^{del} & \text{(14a)} \\ \psi_i^{del} = (1-\epsilon^{del}) \Psi_i^{del} + \epsilon^{ins} \Psi_{n-i}^{ins} & \text{(14b)} \end{cases}$$

where Ψ_i^{ins} and Ψ_i^{del} are the uncorrected SFSs for insertions and deletions, respectively, and are calculated in the same way as (9). As in the SNP case, we only consider cases where the DFE follows a reflected Γ distribution. The shape and scale parameters for insertions and deletions are denoted by a^{ins}, b^{ins}, a^{del}, and b^{del}, respectively.

Parameter estimation:

Let $X^{ins} = (x_1^{ins}, x_2^{ins}, ..., x_{n-1}^{ins})$ and $X^{del} = (x_1^{del}, x_2^{del}, ..., x_{n-1}^{del})$ be the observed SFSs for insertions and deletions, respectively. The log likelihood of the data is calculated as

$$L(\Theta|X^{ins},X^{del}) = \sum_{z\in\{ins,\ del\}} \sum_{i=1}^{n-1} \left(-\psi_i^z + x_i^z \ln(\psi_i^z)\right). \tag{15}$$

Controlling for demography:

Take the continuous model as an example. The augmented SFSs are

$$\begin{cases} \Psi_i^{ins,*} = r_i \Psi_i^{ins} & \text{(16a)} \\ \Psi_i^{del,*} = r_i \Psi_i^{del} & \text{(16b)} \\ \psi_i^{ins,*} = (1-\epsilon^{ins})\Psi_i^{ins,*} + \epsilon^{del}\Psi_{n-i}^{del,*} & \text{(16c)} \\ \psi_i^{del,*} = (1-\epsilon^{del})\Psi_i^{del,*} + \epsilon^{ins}\Psi_{n-i}^{ins,*} & \text{(16d)} \end{cases}$$

As for the neutral reference, we can in principle use any combinations of SNPs, insertions, and deletions collected from putatively neutrally evolving regions. Assume that we have access to both neutral insertions and neutral deletions, and the observed SFSs are denoted by $X^{ins,(0)}$ and $X^{del,(0)}$, respectively. The additional parameters needed to model the neutral variants include $\theta^{ins,(0)}$, $\epsilon^{ins,(0)}$, $\theta^{del,(0)}$, and $\epsilon^{del,(0)}$, which are denoted collectively by $\Theta^{(0)}$. The log likelihood is

$$\begin{aligned} &L(\Theta,\Theta^{(0)},R|X^{ins},X^{del},X^{ins,(0)},X^{del,(0)}) \\ &= L(\Theta,R|X^{ins},X^{del}) + L(\Theta^{(0)},R|X^{ins,(0)},X^{del,(0)}) \end{aligned} \tag{17}$$

where the two terms on the right are calculated using (15) with ψ_i^z replaced by $\psi_i^{z,*}$ ($z \in \{ins,\ del\}$).

Results and Discussion

Simulation results

We evaluate the statistical properties of the new models using computer simulations. Unless stated otherwise, the sample size (n) is 50 and the results are based on 100 replicates. In all cases, we assume the population size is constant and only analyse data from the selected region (see Materials and Methods for justification). For the SNP models, we only present results for the discrete SNP model with $C > 1$ site classes, because both the $C = 1$ case and the continuous model have been analysed before (Glémin *et al.*, 2015; Tataru *et al.*, 2017).

Properties of the discrete SNP model:

First consider a model with $C = 2$ site classes. As can be seen from Table 1, there is information in the SFS for simultaneously estimating all the parameters to a high degree of accuracy. Before discussing more simulation results, it should be pointed out that, when $C > 1$, the order of the site classes is arbitrary. That is, the model considered in Table 1 is equivalent to one with parameters $\theta_1 = 0.01$, $\gamma_1 = -20$, $\epsilon_1 = 0.01$, $\theta_2 = 0.005$, $\gamma_2 = -5$, and $\epsilon_2 = 0.05$. For both cases shown in Table 1, all the MLEs can be sorted such that $\hat{\theta}_1 < \hat{\theta}_2$ and $\hat{\gamma}_1 > \hat{\gamma}_2$. In other words, the MLEs can be assigned unambiguously to site classes according to the order given in the "True value" row. However, if we were to reduce the amount of data, parameter estimates will become more uncertain, and cases such as those with $\hat{\theta}_1 < \hat{\theta}_2$ and $\hat{\gamma}_1 < \hat{\gamma}_2$ will occur, which makes assigning the MLEs to site classes impossible. Thus, presenting mean and standard deviation of the MLEs may give misleading information about the performance of the model.

Table 1. Maximum likelihood estimates (MLEs) of the parameters of discrete SNP models with $C=2$ classes of sites

	m	θ_1	γ_1	ϵ_1	θ_2	γ_2	ϵ_2
True value	–	0.005	-5	0.05	0.01	-20	0.01
Mean (SD) of MLEs	10^6	0.0050 (0.0007)	-5.0 (0.4)	0.051 (0.006)	0.010 (0.001)	-20.2 (1.9)	0.009 (0.006)
Mean (SD) of MLEs	10^5	0.0044 (0.0017)	-4.4 (1.5)	0.042 (0.022)	0.011 (0.001)	-20.0 (5.7)	0.016 (0.014)

NOTE.—Simulated data were generated using the parameter values shown in the "True value" row, with two different region sizes, m. For each parameter combination, 100 samples of size 50 were simulated and analysed to obtain MLEs.

Table 2. Statistical properties of the discrete SNP model

Case	Parameters	m	Percent significant			$\bar{\mu}$				
			Equal ϵ	$\epsilon=0$	$C-1$	True	Full	Equal ϵ	$\epsilon=0$	$C-1$
1	Same as Table 1	10^6	93	100	100	0.0113	0.0114	0.0171	>1	0.0022
2	Same as Table 1	10^5	15	92	100	0.0113	0.0158	0.0204	>1	0.0022
3	See notes below	10^7	3	100	100	0.2204	0.2267	0.2613	>1	0.1755
4	Same as Case 3	2×10^6	0	33	55	0.2204	0.2271	0.2580	>1	0.1768

NOTE.—The parameters used in Case 3 were $\theta_1=0.002$, $\gamma_1=0$, $\epsilon_1=0.05$, $\theta_2=0.006$, $\gamma_2=-5$, $\epsilon_2=0.02$, $\theta_3=0.002$, $\gamma_3=-30$, $\epsilon_3=0.01$, and $n=100$. A large sample size was used for Cases 3 and 4 due to the inclusion of strongly deleterious mutations (i.e., $\gamma_3=-30$). Values under "Percent significant" show how often the full model fitted the data better than the three reduced models (see the main text for more details). The $\bar{\mu}$ (see (18) in Materials and Methods) obtained under the $\epsilon=0$ model are large because ignoring polarisation error results in the inference of a site class with a strongly positive γ.

In light of the above discussion, we investigate the statistical properties of the model using two alternative methods. First, we compare the full model to the following reduced models using the χ^2 test: "Equal ϵ" (all site share the same polarisation error rate), "$\epsilon=0$" (no polarisation error), and "$C-1$" (a model with $C-1$ site classes, where C is the true number of site classes). Second, we assess how well these various models predict the average fixation probability $\bar{\mu}$ (see (18) in Materials and Methods), which is essential for estimating the prevalence of adaptive substitutions (i.e., α and ω_a).

Considering the two pairs of cases in Table 2, and focusing on the data presented under "Percent significant", we make the following observations. First, as the amount of data reduces, the ability of the model to infer separate ϵ for different site classes drops more rapidly than its ability to detect the existence of either polarisation error or more than one site class. This suggests that estimating heterogeneity in ϵ may be challenging. Considering all four cases, it appears that the tests for detecting the presence of polarisation error (i.e., the full model versus "$\epsilon=0$") and for detecting the existence of more site classes (i.e., the full model versus "$C-1$") are more powerful, especially the latter. It should be noted that the likelihood surface appears to be rather flat when $C=3$ such that different parameter combinations may produce very similar log likelihoods. This is particularly evident when the amount of data is limited (Case 3 versus Case 4), leading to a reduction in power of the tests. A similar observation was made by Keightley and Eyre-Walker (2010), who also showed that it can be partly alleviated by increasing the sample size. Nonetheless there may well be a limit as to how many site classes can be included. This identifiability problem is analogous to that

Table 3. MLEs of the parameters of several INDEL models

Model	m	Parameters								
Discrete	2×10^6	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}		
		True	0.0005	-5	0.02	0.001	-15	0.02		
		Mean MLE	0.00050	-5.0	0.021	0.0010	-15.0	0.020		
Continuous	2×10^7	Name	θ^{ins}	a^{ins}	b^{ins}	ϵ^{ins}	θ^{del}	a^{del}	b^{del}	ϵ^{del}
		True	0.0005	0.5	10	0.08	0.001	0.25	50	0.04
		Mean MLE	0.00050	0.51	10.4	0.080	0.0010	0.251	51.2	0.040
Continuous	2×10^6	Name	θ^{ins}	a^{ins}	b^{ins}	ϵ^{ins}	θ^{del}	a^{del}	b^{del}	ϵ^{del}
		True	0.0005	0.5	10	0.08	0.001	0.25	50	0.04
		Mean MLE	0.00054	0.51	144.7	0.082	0.0010	0.253	93.2	0.041

discussed extensively in the context of using SNP-based methods for estimating past demographic changes (e.g., Myers *et al.*, 2008).

Interestingly, the reduced model "Equal ϵ" makes worse predictions of $\bar{\mu}$ than the full model in all cases presented in Table 2, even when the full model does not normally provide a better fit to the data (Cases 2 and 4). The same applies to the other two reduced models. Thus, despite the statistical difficulties discussed above, fitting the full model to the data may be important for obtaining accurate estimates of α and ω_a.

Properties of the INDEL models:

Table 3 contains simulation results based on a discrete model (with $C^{ins}=C^{del}=1$) and two continuous models (differing from each other in terms of the size of the focal region m). The mutation rates are about 10 times lower than those used in the SNP cases (Tables 1 and 2), and polarisation error rates are about 2 times higher. These choices are to reflect the fact that INDELs are generally less prevalent than SNPs, and are potentially more difficult to polarise. As can be seen, with a reasonable amount of data, all the parameters can be reliably estimated. Comparing the two continuous models, we notice that, with limited data, the scale parameter b of the Γ distribution may be overestimated, but estimates of the shape parameter a and the polarisation error rate remain unbiased.

The true values of $\bar{\mu}^{ins}$ and $\bar{\mu}^{del}$ for the discrete model are 0.0339 and 4.59×10^{-6}, respectively. The mean (SD) of the estimates is 0.0345 (0.0055) for $\bar{\mu}^{ins}$, and 5.27×10^{-6} (2.91×10^{-6}) for $\bar{\mu}^{del}$. Thus, the true values are well within the observed ranges of variability. The true values of $\bar{\mu}^{ins}$ and $\bar{\mu}^{del}$ for the two continuous cases are 0.384 and 0.429, respectively. The mean (SD) of the estimates for the case with more data is 0.382 (0.012) for $\bar{\mu}^{ins}$ and 0.429 (0.008) for $\bar{\mu}^{del}$. Encouragingly, for the continuous case with less data, despite the tendency to overestimate the scale parameter, estimates of the average fixation probabilities are still highly accurate: 0.388 (0.050) for $\bar{\mu}^{ins}$ and 0.418 (0.028) for $\bar{\mu}^{del}$, suggesting that the reliability of estimates of α and ω_a is unlikely to be compromised.

Table 4. Summary statistics for the INDEL and SNP data

Data	Type	Diversity (π)	Tajima's D
INDELs	CDS	5.20×10^{-5}	-1.208
	Frameshift	2.06×10^{-5}	-1.253
	Non-frameshift	3.14×10^{-5}	-1.177
	Intron	0.0016	-0.729
	Intergenic	0.0017	-0.704
	Non-coding	0.0017	-0.718
SNPs	Nonsense	5.83×10^{-6}	-1.510
	0-fold degenerate sites	0.0016	-0.868
	4-fold degenerate sites	0.0165	-0.210

Application to *D. melanogaster* data

A summary of the data

Using the variant calling pipeline detailed in Materials and Methods, a total of 370,217 INDELs ($\leq$ 50bp) and 1,789,367 SNPs were identified from the 17 Rwandan individuals. Our analysis primarily focuses on INDELs because SNPs have been analysed extensively before (Eyre-Walker and Keightley, 2009; Keightley and Eyre-Walker, 2007; Schneider *et al.*, 2011). Similar to previous reports (e.g., Ptak and Petrov, 2002), smaller INDELs are more prevalent than larger ones (Figure S1). INDEL diversity is about 30 times lower in protein-coding (CDS) regions than in either intronic or intergenic regions (Table 4). Additionally, frameshift INDELs are rarer than non-frameshift ones (Table 4; supplementary Figure S1). Interestingly, nonsense mutations are somewhat rarer than frameshift INDELs, an observation also made by Leushkin *et al.* (2013). These results indicate strong purifying selection against INDELs in protein-coding regions. INDEL diversity patterns appear to be similar between intronic and intergenic regions. They are combined and referred to as non-coding INDELs in what follows to increase statistical power.

Comparing between INDELs and SNPs, we notice that INDEL diversity in non-coding regions is about 10 times lower than π_4 (4-fold site diversity; Table 4), consistent with the fact that the INDEL mutation rate is lower than the point mutation rate (Haag-Liautard *et al.*, 2007; Schrider *et al.*, 2013). However, Tajima's D calculated on non-coding INDELs is more negative than that calculated on 4-fold sites (Table 4), probably reflecting the fact that many non-coding DNA in the *D. melanogaster* genome are under selection (Andolfatto, 2005). Furthermore, π_0 (0-fold site diversity; Table 4) is only about 10 times smaller than π_4. This level of reduction is much smaller than the 30-fold difference observed between CDS and non-coding INDELs. This suggests that, in protein-coding regions, INDEL mutations are under much stronger purifying selection than 0-fold mutations, which is consistent with the more negative Tajima's D value calculated on CDS INDELs (Table 4).

To further investigate the data, we calculated d_N, substitution rate at nonsynonymous sites, using PAML and the reference genomes of *D. simulans* and *D. yakuba* (see Materials and Methods). The genes were then divided into 20 equal-sized bins. For each bin, we calculated average π_0 and π_{INDEL}. Both statistics decrease as d_N decreases (Figure S2), consistent with the

Table 5. Results based on the best-fitting models for INDELs in the CDS regions of the *D. melanogaster* genome.

Neutral ref/DFE/mutation rate	Parameters for CDS INDELs							α
Noncoding INDELs	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}	83.7%
Discrete $C=2$	MLE	1.8×10^{-5}	1.98	0.023	5.3×10^{-5}	-1.69	0.016	
Uniform mutation rate	Name	θ_2^{ins}	γ_2^{ins}	ϵ_2^{ins}	θ_2^{del}	γ_2^{del}	ϵ_2^{del}	
	MLE	7.2×10^{-4}	-1566.4	3.6×10^{-5}	0.0011	-642.5	1.6×10^{-5}	
4-fold degenerate sites	Name	θ_1^{ins}	γ_1^{ins}	ϵ_1^{ins}	θ_1^{del}	γ_1^{del}	ϵ_1^{del}	71.5%
Discrete $C=2$	MLE	1.6×10^{-5}	-1.31	0.0092	4.9×10^{-5}	-3.77	0.0082	
Fixed mutation ratios	Name	θ_2^{ins}	γ_2^{ins}	ϵ_2^{ins}	θ_2^{del}	γ_2^{del}	ϵ_2^{del}	
	MLE	1.9×10^{-4}	-284.1	1.2×10^{-4}	0.0010	-454.8	6.2×10^{-5}	

NOTE.—The DFE for polymorphic INDELs in the CDS regions were inferred using either non-coding INDELs or 4-fold sites as the neutral reference. A series of different DFEs were fitted to the data, and the best-fitting models presented above were determined by using the Akaike information criterion (AIC) (see supplementary Tables S1 and S3). When non-coding INDELs were used as the neutral reference, α was estimated using INDEL divergence in noncoding regions. When 4-fold sites were used as the neutral reference, the mutation rate ratio between SNPs and INDELs, and that between deletions and insertions, were fixed at values obtained from a mutation accumulation experiment (Schrider *et al.*, 2013). α was estimated using a method based on divergence in the 8–30bp region of short introns < 66bp long (see the main text).

expectation that mutations are on average more deleterious in more conserved genes (Jackson *et al.*, 2015). The results in this and the preceding paragraphs suggest that our INDEL dataset is of high quality.

Inferring the DFE and α using non-coding INDELs as the neutral reference

To infer the DFE for INDELs in CDS regions, we used non-coding INDELs as the neutral reference. Following previous efforts in estimating the DFE for SNPs (Eyre-Walker and Keightley, 2009; Galtier, 2016; Keightley and Eyre-Walker, 2007; Schneider *et al.*, 2011; Tataru *et al.*, 2017), we also assumed that the mutation rate towards insertions and deletions, respectively, were the same between the neutral and selected regions. The best-fitting DFE is one with $C=2$ classes of selected sites (Table 5 and supplementary Table S1). The MLEs of γ suggest that polymorphic INDELs are either nearly neutral or are so strongly deleterious that they contribute little to polymorphism. This seems to be consistent with the 30-fold difference in INDEL diversity level between CDS and non-coding regions, which is more substantial than the 10-fold difference between 0-fold and 4-fold sites (Table 4). Fitting the data to a discrete model with $C=3$ classes of sites also reveals a bimodal DFE, suggesting that the conclusion is robust (supplementary Table S1). With a larger sample containing hundreds or even thousands of alleles, and by fitting a DFE with more site classes, it should be possible to obtain further details of the relative frequencies and fitness effects of strongly selected variants, which tend not to segregate in our current sample of size 17. However, this additional information about the strongly selected end of the DFE is unlikely to affect our estimation of α (see below) because these variants make effectively no contribution to divergence.

To better understand the effects of length, we separated the INDELs in CDS regions into the following length categories: 1bp, 2bp, 3bp, frameshifting (≥4bp), and non-frameshifting (≥6bp). We analysed the data in each category separately. As above, non-coding INDELs with the same length were used as the neutral reference and the mutation was assumed to be constant across neutral and selected sites. Considering the dearth of variants, we only fitted a DFE with $C = 1$ class of selected sites. Viewing the γ in this model as the "average" selection coefficient, frameshift INDELs are consistently more deleterious than non-frameshift INDELs (supplementary Figure S3). Consistent with a prevous study (Leushkin *et al.*, 2013), there is no obvious evidence that longer INDELs are under stronger selection.

Using the best-fitting DFE (Table 5), the proportion of INDEL substitutions in the CDS regions fixed by positive selection in the *D. melanogaster* lineage, α, is 83.7% (100% for insertions and 81.8% for deletions). These α estimates are comparable to previous estimates for SNP substitutions in CDS regions (Andolfatto *et al.*, 2011; Schneider *et al.*, 2011).

As mentioned above, some non-coding INDELs are probably non-neutral, as suggested by the negative Tajima's D value (Table 4). Our use of these variants as the neutral reference are for several practical reasons. Although using INDELs in "dead-on-arrival" transposable elements as neutral reference may be preferable (Petrov, 2002), calling variants from repetitive regions using short-read data is highly prone to error (Li, 2014). Using data from the 8-30bp region of short introns $\leq$ 65bp, which are also putatively neutral (Parsch *et al.*, 2010), is also problematic because of evidence for selection maintaining intron size (Leushkin *et al.*, 2013; Parsch, 2003; Ptak and Petrov, 2002). Note that Tajima's D is more negative for INDELs in CDS regions than for those in non-coding regions, suggesting that the latter are probably under weaker purifying selection (Table 4). If this is the case, our method tends to underestimate the strength of purifying selection on INDELs in CDS regions, as suggested by the simulation results presented in supplementary Table S2. This should lead to an overestimation of $\bar{\mu}$, the average fixation rate (Eq. (18)), which should in turn put a downward pressure on the estimation of α (Eq. (19)). However, biases in α also depend on the way selection on non-coding INDELs alters divergence. For example, if fixations of beneficial non-coding INDELs are so common that d_S is greater than the divergence level expected under neutral evolution, then this combined with the overestimation of $\bar{\mu}$ can lead to a substantial underestimation of α. In contrast, if most non-coding INDELs are selected against and d_S is much smaller than the neutral expectation, it may offset the effect caused by the overestimation of $\bar{\mu}$ and result in an overestimation of α.

Inferring the DFE and α using 4-fold degenerate sites as the neutral reference

To check the robustness of our results, we conducted a second set of analyses without using non-coding INDELs. We extended our model such that it can infer the DFE for INDELs in CDS regions using 4-fold sites as the neutral reference. We chose 4-fold sites instead of the 8-30bp region of short introns $\leq$ 65bp because 4-fold sites are probably not under ongoing selection on codon usage in *D. melanogaster*, and are similar to short introns in multiple aspects of polymorphism patterns (Jackson *et al.*, 2017). Considering the parameter richness of the models, using 4-fold SNPs as the neutral reference should help statistical inference because they are much more numerous than short-intron SNPs.

We used the following approach to obtain neutral divergence for INDELs along the *D. melanogaster* lineage. The nucleotide divergence in the 8-30bp region of short introns $\leq$ 65bp is 0.0674 (B. Jackson personal communication). In a mutation accumulation experiment (Schrider *et al.*, 2013), it was found that the rate to point mutations is 12.2 times higher than that to short INDELs, and that the rate to deletions is 5 times higher than that to insertions (averaging across the two genetic backgrounds considered therein). Thus, an estimate of neutral INDEL divergence can be obtained as $0.0674/12.2 = 0.0055$, and the corresponding estimates for insertions and deletions are 9.2×10^{-4} and 0.0046, respectively.

Due to the use of 4-fold sites as the neutral reference, it is no longer appropriate to assume that the mutation rate is the same between the selected and neutral regions. Given the evidence that the DFE for INDELs probably features a class of strongly deleterious mutations that make little contribution to polymorphism, allowing the selected and neutral regions to have their separate mutation rates is likely to cause the model to underestimate both the mutation rate in the selected region and strength of purifying selection, as confirmed by simulation results presented in supplementary Table S3. An underestimation of the strength of purifying selection is likely to cause an underestimation of α. We observed this in our dataset – α for all INDELs obtained from the best-fitting DFE for this analysis (supplementary Table S4) is only 21.7%, much smaller than the value of 83.7% when non-coding INDELs were used as the neutral reference (Table 5).

To resolve the above problem, we again made use of the information reported in the aforementioned mutation accumulation experiment (Schrider *et al.*, 2013). Specifically, we further extended our model, so that the mutation rate ratio between SNPs and INDELs, and that between deletions and insertions, were fixed at 12.2 and 5, respectively. As shown in Table 5 (see also supplementary Table S5), the best-fitting DFE has $C = 2$ class of sites, with one under weak selection, and the other being strongly deleterious. The α estimates for all INDELs, insertions and deletions are, respectively, 71.5%, 59.7%, and 81.3%.

To make sure that the above results are not dependent on our use of the mutation rate ratios estimated by Schrider *et al.* (2013), we repeated the analysis using ratios obtained by either Petrov and Hartl (1998) (SNP/INDEL

= 6.9 and deletion/insertion = 8.7) or Haag-Liautard *et al.* (2007) (SNP/INDEL = 4.2 and deletion/insertion = 3.0) (supplementary Table S6). In both cases, the best-fitting DFE has $C = 2$ classes of selected sites, under weak and strong selection, respectively (supplementary Tables S7 and S8). Furthermore, estimates of the strength of purifying selection acting on sites in the weakly selected class are almost identical regardless of the choice of mutation rate ratios (supplementary Table S9). Thus, unsurprisingly, all three analyses also produce very similar α estimates (supplementary Table S9). Overall, these results are consistent with those based on non-coding INDELs and suggest that a substantial fraction of INDEL substitutions were fixed by positive selection.

Materials and Methods

Numerical details

We used numerical routines provided by the GNU Scientific Library (GSL; `https://www.gnu.org/software/gsl/`) to perform the integration in (3) numerically. For the continuous model (e.g., (9)), the integral was evaluated using Gaussian quadrature, which was implemented based on a routine included in the R package `statmod` (`https://cran.r-project.org/web/packages/statmod/index.html`). Maximum likelihood estimates of the model parameters were obtained by both gradient-based and derivative-free optimization algorithms implemented in the `NLopt` package (`http://ab-initio.mit.edu/wiki/index.php/NLopt`). To ensure the global maximum was found, we initialised the search algorithm using multiple randomly selected starting points.

Simulations

We performed parameter estimation using our program, `anavar`, on random samples simulated using Mathematica (`http://www.wolfram.com/`). Because the generation of simulated data is separate from the numerical routines we used to implement `anavar`, this set-up can help verify the numerical robustness of `anavar`. Note that, in all simulations, we only used the models to analyse variants from selected regions because we wanted to find out how much information we could obtain by analysing them alone. Including neutral variants, as routinely done in real data analysis, may help to increase the accuracy of parameter estimation. So our choice should give us a rather conservative assessment of the methods' performance.

In addition to testing whether the data contained enough information for all the parameters to be estimated, we also assessed how well a model could predict the average fixation rate, $\bar{\mu}$ (expressed in units of $2N_e$ generations). As an example, if nonsynonymous polymorphism data are fitted to the discrete SNP model, $\bar{\mu}$ can be estimated as

$$\bar{\mu} = \frac{1}{\hat{\bar{\theta}}} \sum_{c=1}^{C} \frac{\hat{\theta}_c \hat{\gamma}_c}{1 - e^{-\hat{\gamma}_c}} \qquad (18)$$

where $\hat{Z}$ signifies the MLE of parameter Z and θ is defined by (1). Understanding the ability to accurately estimate $\bar{\mu}$ is important because it is needed for estimating α, the proportion of substitutions fixed by positive selection, which can be written as,

$$\alpha=\frac{d_N-d_S\bar{\mu}}{d_N} \tag{19}$$

where d_N and d_S are the numbers of selected (e.g., nonsynonymous) and neutral (e.g., synonymous) substitutions per site, respectively (Eyre-Walker and Keightley, 2009).

We did not generate simulated data from models with demographic changes and selection at linked sites because the effectiveness of the method of Eyre-Walker *et al.* (2006) in controlling for these confounding factors have been studied extensively (Eyre-Walker *et al.*, 2006; Galtier, 2016; Glémin *et al.*, 2015; Jackson *et al.*, 2017; Muyle *et al.*, 2011; Tataru *et al.*, 2017).

The *Drosophila melanogaster* dataset

This dataset consisted of 17 Rwandan individuals as described in Jackson *et al.* (2015, 2017) and made available by the *Drosophila* Population Genomics Project (Pool *et al.*, 2012).

Variant calling:

INDEL realigned BAM files were obtained from Jackson *et al.* (2017). Initial genotype calling was performed with the HaplotypeCaller and GenotypeGVCF (with the `-includeNonVariantSites` flag to output genotype calls at both variant and non-variant positions) tools from GATK 3.7 (DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013). Variant quality score recalibration (VQSR) requires one 'truth set' for SNPs and one for INDELs. To generate the truth sets, we intersected the raw variants called from GATK with variants called from SAMtools (version 1.2) (Li *et al.*, 2009). The consensus data was further filtered using the GATK best practice hard filters (for SNPs: QD < 2.0, MQ < 40.0, FS > 60.0, SOR > 3.0, MQRankSum < -12.5, ReadPosRankSum < -8.0; for INDELs: QD < 2.0, ReadPosRankSum < -20.0, FS > 200.0, SOR > 10.0; see `https://software.broadinstitute.org/gatk/guide/article?id=3225`). Variants with coverage more than twice, or less than half, the mean coverage of 20X were excluded, along with variants falling into regions identified by `RepeatMasker` (`http://www.repeatmasker.org`). Multiallelic sites were excluded along with SNPs falling within INDELs and INDELs greater than 50bp. We ran VQSR separately for SNPs and INDELs, retaining variants that fell within the 95% tranche cut-off as in Jackson *et al.* (2017). The passing variants were then re-filtered as above with the exception of the GATK hard filters which were not reapplied.

Multi-species alignments and polarisation:

Multi-species alignments were generated between *D. melanogaster* (v5.34), *D. simulans* (Hu *et al.*,

2013) and *D. yakuba* (v1.3) using *D. melanogaster* as reference. Firstly pairwise alignments were created using LASTZ (Harris, 2007). These were then chained and netted using axtChain and chainNet, respectively (Kent *et al.*, 2003). Single coverage was ensured for the reference genome using single_cov2.v11 from the MULTIZ package (Blanchette *et al.*, 2004) and the pairwise alignments were aligned with MULTIZ.

Variants were polarised using the whole genome multi-species alignment and a parsimony approach, where either the alternate or the reference allele had to be supported by all outgroups in the the alignment to be considered ancestral. The site-frequency spectra for insertions and deletions in different genomic regions are presented in supplementary Figure S4.

Annotation:

Variants were annotated as either intronic, intergenic or CDS using the *D. melanogaster* GFF annotation file (version 5.34, available from: `ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.34_FB2011_02/gff/`). Fourfold degenerate and zerofold degenerate SNPs in CDS regions were annotated using coordinates obtained from the *D. melanogaster* CDS fasta sequences (version 5.34, available from: `ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.34_FB2011_02/fasta/dmel-all-CDS-r5.34.fasta.gz`).

Summary statistics:

Nucleotide diversity (π) (Tajima, 1983), Watterson's θ (Watterson, 1975) and Tajima's D (Tajima, 1989) were calculated for variants in non-coding (intronic and intergenic) and coding regions, as well as for 0-fold and 4-fold degenerate SNPs. The numbers of callable sites used to obtain per-site estimates was taken to be the number of sites in each region that were called in the "all sites" VCF file and passed the filters described previously. Additionally for polarised variants the number of callable sites was reduced to those that could be polarised by our parsimony approach.

To obtain rates of divergence at nonsynonymous and synonymous sites, denoted by d_N and d_S, CDS regions were extracted from the multi-species alignment using the coordinates from the *D. melanogaster* CDS fasta alignment file. CDS alignments were removed if they were not in frame, did not start with a start codon, did not end with a stop codon or contained premature stop codons. Additionally any codons with missing data were dropped. For each gene we retained only the longest transcript. This data was then analysed using codeml in PAML (Yang, 2007) with a one ratio model to obtain d_N and d_S.

Supplementary Material

The new models have been implemented in a user-friendly package `anavar`, which is freely available at `http://zeng-lab.group.shef.ac.`

uk. In addition to the models developed herein, anavar also contains implementations of several other widely-used models for estimating the DFE (i.e., Eyre-Walker *et al.*, 2006) and for studying GC-biased gene conversion (gBGC) (i.e., Glémin *et al.*, 2015). All scripts used for the anavar simulation analyses are available at https://github.com/henryjuho/anavar_simulations. Additionally, all scripts used in the *D. melanogaster* analyses can be found at https://github.com/henryjuho/drosophila_indels.

Acknowledgments

We thank Ben Jackson and Jose Campos for providing valuable assistance in the analysis of the *Drosophila* data. This work was supported by a PhD studentship funded by the Department of Animal and Plant Sciences, University of Sheffield, to H.J.B. Support was also provided by the Natural Environment Research Council via a research grant awarded to K.Z. (NE/L005328/1).

References

Ananda, G., Walsh, E., Jacob, K. D., Krasilnikova, M., Eckert, K. A., Chiaromonte, F., and Makova, K. D. 2013. Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome. *Genome Biol Evol*, 5(3): 606–20.

Andolfatto, P. 2005. Adaptive evolution of non-coding dna in drosophila. *Nature*, 437(7062): 1149–52.

Andolfatto, P., Wong, K. M., and Bachtrog, D. 2011. Effective population size and the efficacy of selection on the x chromosomes of two closely related drosophila species. *Genome Biol Evol*, 3: 114–28.

Besenbacher, S., Liu, S., Izarzugaza, J. M. G., Grove, J., Belling, K., Bork-Jensen, J., Huang, S., Als, T. D., Li, S., Yadav, R., Rubio-García, A., Lescai, F., Demontis, D., Rao, J., Ye, W., Mailund, T., Friborg, R. M., Pedersen, C. N. S., Xu, R., Sun, J., Liu, H., Wang, O., Cheng, X., Flores, D., Rydza, E., Rapacki, K., Damm Sørensen, J., Chmura, P., Westergaard, D., Dworzynski, P., Sørensen, T. I. A., Lund, O., Hansen, T., Xu, X., Li, N., Bolund, L., Pedersen, O., Eiberg, H., Krogh, A., Børglum, A. D., Brunak, S., Kristiansen, K., Schierup, M. H., Wang, J., Gupta, R., Villesen, P., and Rasmussen, S. 2015. Novel variation and de novo mutation rates in population-wide de novo assembled danish trios. *Nat Commun*, 6: 5969.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D., and Miller, W. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, 14(4): 708–15.

Bustamante, C. D., Wakeley, J., Sawyer, S., and Hartl, D. L. 2001. Directional selection and the site-frequency spectrum. *Genetics*, 159(4): 1779–88.

Corcoran, P., Gossmann, T. I., Barton, H. J., Great Tit HapMap Consortium, Slate, J., and Zeng, K. 2017. Determinants of the efficacy of natural selection on coding and noncoding variability in two passerine species. *Genome Biol Evol*, 9(11): 2987–3007.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. 2011. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nat Genet*, 43(5): 491–8.

Earl, D., Nguyen, N., Hickey, G., Harris, R. S., Fitzgerald, S., Beal, K., Seledtsov, I., Molodtsov, V., Raney, B. J., Clawson, H., Kim, J., Kemena, C., Chang, J.-M., Erb, I., Poliakov, A., Hou, M., Herrero, J., Kent, W. J., Solovyev, V., Darling, A. E., Ma, J., Notredame, C., Brudno, M., Dubchak, I., Haussler, D., and Paten, B. 2014. Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Res*, 24(12): 2077–89.

Eyre-Walker, A. and Keightley, P. D. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet*, 8(8): 610–8.

Eyre-Walker, A. and Keightley, P. D. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol*, 26(9): 2097–108.

Eyre-Walker, A., Woolfit, M., and Phelps, T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, 173(2): 891–900.

Galtier, N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS Genet*, 12(1): e1005774.

Glémin, S., Arndt, P. F., Messer, P. W., Petrov, D., Galtier, N., and Duret, L. 2015. Quantification of gc-biased gene conversion in the human genome. *Genome Res.*, 25(8): 1215–28.

Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D. L., Houle, D., Charlesworth, B., and Keightley, P. D. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in drosophila. *Nature*, 445(7123): 82–5.

Harris, R. S. 2007. *Improved pairwise alignment of genomic DNA*. The Pennsylvania State University.

Hartfield, M. and Keightley, P. D. 2012. Current hypotheses for the evolution of sex and recombination. *Integr Zool*, 7(2): 192–209.

Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol Biol Evol*, 24(8): 1792–800.

Hu, T. T., Eisen, M. B., Thornton, K. R., and Andolfatto, P. 2013. A second-generation assembly of the drosophila simulans genome provides new insights into patterns of lineage-specific divergence. *Genome Res.*, 23(1): 89–98.

Jackson, B. C., Campos, J. L., and Zeng, K. 2015. The effects of purifying selection on patterns of genetic differentiation between drosophila melanogaster populations. *Heredity (Edinb)*, 114(2): 163–74.

Jackson, B. C., Campos, J. L., Haddrill, P. R., Charlesworth, B., and Zeng, K. 2017. Variation in the intensity of selection on codon bias over time causes contrasting patterns of base composition evolution in drosophila. *Genome Biol Evol*, 9(1): 102–123.

Keightley, P. D. and Eyre-Walker, A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, 177(4): 2251–61.

Keightley, P. D. and Eyre-Walker, A. 2010. What can we learn about the distribution of fitness effects of new mutations from dna sequence data? *Philos Trans R Soc Lond B Biol Sci*, 365(1544): 1187–93.

Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., and Blaxter, M. L. 2009. Analysis of the genome sequences of three drosophila melanogaster spontaneous mutation accumulation lines. *Genome Res*, 19(7): 1195–201.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A*, 100(20): 11484–9.

Kim, B. Y., Huber, C. D., and Lohmueller, K. E. 2017. Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics*, 206(1): 345–361.

Kousathanas, A. and Keightley, P. D. 2013. A comparison of models to infer the distribution of fitness effects of new mutations. *Genetics*, 193(4): 1197–208.

Kvikstad, E. M. and Duret, L. 2014. Strong heterogeneity in mutation rate causes misleading hallmarks of natural selection on indel mutations in the human genome. *Mol Biol Evol*, 31(1): 23–36.

Leushkin, E. V. and Bazykin, G. A. 2013. Short indels are subject to insertion-biased gene conversion. *Evolution*, 67(9): 2604–13.

Leushkin, E. V., Bazykin, G. A., and Kondrashov, A. S. 2013. Strong mutational bias toward deletions in the drosophila melanogaster genome is compensated by selection. *Genome Biol Evol*, 5(3): 514–24.

Li, H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20): 2843–51.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup 2009. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16): 2078–9.

Montgomery, S. B., Goode, D. L., Kvikstad, E., Albers, C. A., Zhang, Z. D., Mu, X. J., Ananda, G., Howie, B., Karczewski, K. J., Smith, K. S., Anaya, V., Richardson, R., Davis, J., 1000 Genomes Project Consortium, MacArthur, D. G., Sidow, A., Duret, L., Gerstein, M., Makova, K. D., Marchini, J., McVean, G., and Lunter, G. 2013. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res*, 23(5): 749–61.

Muyle, A., Serres-Giardi, L., Ressayre, A., Escobar, J., and Glémin, S. 2011. Gc-biased gene conversion and selection affect gc content in the oryza genus (rice). *Mol Biol Evol*, 28(9): 2695–706.

Myers, S., Fefferman, C., and Patterson, N. 2008. Can one learn history from the allelic spectrum? *Theor Popul Biol*, 73(3): 342–8.

Parsch, J. 2003. Selective constraints on intron evolution in drosophila. *Genetics*, 165(4): 1843–51.

Parsch, J., Novozhilov, S., Saminadin-Peter, S. S., Wong, K. M., and Andolfatto, P. 2010. On the utility of short intron sequences as a reference for the detection of positive and negative selection in drosophila. *Mol Biol Evol*, 27(6): 1226–34.

Petrov, D. A. 2002. Dna loss and evolution of genome size in drosophila. *Genetica*, 115(1): 81–91.

Petrov, D. A. and Hartl, D. L. 1998. High rate of dna loss in the drosophila melanogaster and drosophila virilis species groups. *Mol Biol Evol*, 15(3): 293–302.

Pool, J. E., Corbett-Detig, R. B., Sugino, R. P., Stevens, K. A., Cardeno, C. M., Crepeau, M. W., Duchen, P., Emerson, J. J., Saelao, P., Begun, D. J., and Langley, C. H. 2012. Population genomics of sub-saharan drosophila melanogaster: African diversity and non-african admixture. *PLoS Genet*, 8(12): e1003080.

Ptak, S. E. and Petrov, D. A. 2002. How intron splicing affects the deletion and insertion profile in drosophila melanogaster. *Genetics*, 162(3): 1233–44.

Sawyer, S. A. and Hartl, D. L. 1992. Population genetics of polymorphism and divergence. *Genetics*, 132(4): 1161–76.

Schneider, A., Charlesworth, B., Eyre-Walker, A., and Keightley, P. D. 2011. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics*, 189(4): 1427–37.

Schrider, D. R., Houle, D., Lynch, M., and Hahn, M. W. 2013. Rates and genomic consequences of spontaneous mutational events in drosophila melanogaster. *Genetics*, 194(4): 937–54.

Tajima, F. 1983. Evolutionary relationship of dna sequences in finite populations. *Genetics*, 105(2): 437–60.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3): 585–95.

Tataru, P., Mollion, M., Glémin, S., and Bataillon, T. 2017. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics*, 207(3): 1103–1119.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and

DePristo, M. A. 2013. From fastq data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics*, 43: 11.10.1–33.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, 7(2): 256–76.

Yang, S., Wang, L., Huang, J., Zhang, X., Yuan, Y., Chen, J.-Q., Hurst, L. D., and Tian, D. 2015. Parent-progeny sequencing indicates higher mutation rates in heterozygotes. *Nature*, 523(7561): 463–7.

Yang, Z. 2007. Paml 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24(8): 1586–91.

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., and Lander, E. S. 2014. Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A*, 111(4): E455–64.