From drift to draft: How much do beneficial mutations actually contribute to predictions of Ohta's slightly deleterious model of molecular evolution?

Jun Chen[*],[†] **Sylvain Glémin**[*],[§], **Martin Lascoux**[*],[‡]

[*]Program in Plant Ecology and Evolution, Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, 75236 Uppsala, Sweden

§ Université de Rennes, CNRS, ECOBIO [Ecosystèmes, Biodiversité, Evolution] - UMR 6553, F-35000 Rennes, France

†Present address: College of Life Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China

‡Author for correspondence: Martin.Lascoux@ebc.uu.se

Abstract

Since its inception in 1973 the slightly deleterious model of molecular evolution, aka the Nearly Neutral Theory of molecular evolution, remains a central model to explain the main patterns of DNA polymorphism in natural populations. This is not to say that the quantitative fit to data is perfect. In a recent study CASTELLANO *et al.* (2018) used polymorphism data from *D. melanogaster* to test whether, as predicted by the Nearly Neutral Theory, the proportion of effectively neutral mutations depends on the effective population size (N_e). They showed that a nearly neutral model simply scaling with N_e variation across the genome could not explain alone the data but that consideration of linked positive selection improves the fit between observations and predictions. In the present article we extended their work in two main directions. First, we confirmed the observed pattern on a set of 59 species, including high quality genomic data from 11 animal and plant species with different mating systems and effective population sizes, hence levels of linked selection. Second, for the 11 species with high quality genomic data we also estimated the full Distribution of Fitness Effects (DFE) of mutations, and not solely the DFE of deleterious mutations. Both N_e and beneficial mutations contributed to the relationship between the proportion of effectively neutral mutations and local N_e across the genome. In conclusion, the predictions of the slightly deleterious model of molecular evolution hold well for species with small effective population size. But for species with large N_e the fit is improved by incorporating linked positive selection to the model.

Keywords: Nearly Neutral Theory, Distribution of Fitness Effects, beneficial mutations, linked selection

Introduction

The year 2018 saw the celebration of the 50th anniversary of the Neutral Theory of molecular evolution (called simply the Neutral Theory thereafter). At 50 years of age, the Neutral Theory is still shrouded in controversies, some pronouncing it dead and overwhelmingly rejected by facts (Kern and Hahn 2018) while others see it as very much alive and kicking (NEI *et al.* 2010, JENSEN *et al.* 2019). As a quick glance at major textbooks in population genetics and at the literature would suggest, it seems fair to say that the Neutral Theory is certainly not totally dead. Even if it undoubtedly did lose some of its initial appeal it continues to play a central role in population genetics, a position well summarized by Kreitman (1996) in his spirited essay "The neutral theory is dead. Long live the neutral theory". Shortcomings of the Neutral Theory were already noted in the 1970s and the Neutral Theory has itself evolved. Indeed, its inadequacy to fully explain the data, in particular Lewontin's paradox (Lewontin 1974; Corbett-Detig et al. 2015), was already noted in 1973, leading Tomoko Ohta (1973) to propose the Nearly Neutral Theory of molecular evolution. In contrast to the Neutral Theory where most mutations are assumed to be neutral or strongly deleterious, the Nearly Neutral Theory assigns much more prominence to the contribution to standing polymorphism of mutations that are weakly selected and effectively neutral (Ohta 1992; Ohta and Gillespie 1996). Weakly selected mutations can be slightly deleterious or slightly beneficial, but as noted by Kreitman (1996) the best developed of the weak selection models primarily consider slightly deleterious mutations and was therefore christened by him "the slightly deleterious model". This is the model that we will be testing in most of the present paper.

Like the Neutral Theory, however, the Nearly Neutral Theory still assumes that "only a minute fraction of DNA changes in evolution are adaptive in nature" (Kimura 1983). Under this view, polymorphism is thought to be mostly unaffected by positive selection, except around the few recently selected beneficial alleles (selective sweep). This was already at variance with the view put forward by Gillespie (e.g. Gillespie 2004) that assigned a greater role to linked positive selection in shaping polymorphism (see also Corbett-Detig *et al.* 2015) and is in even stronger contrast with the claim by Kern and Hahn (2018) that "natural selection has played the predominant role in shaping within- and between-species genetic variation" and that "the ubiquity of adaptive variation both within and between species" leads to the rejection of the universality of the Neutral Theory. In a far more nuanced assessment of the Neutral Theory and its contribution, Jensen *et al.* (2018) argued that the effects of linked selection could readily be incorporated in the Nearly Neutral framework. The core of the dispute, either today or in the early days of the Nearly Neutral Theory, is about the degree to which each category of mutations contributed directly and indirectly to genetic variation within- and between-species.

A core prediction of the Nearly Neutral Theory is that the fraction of mutations affected by selection depends on N_e (Ohta 1973). N_e can vary among species but also within a genome because of linked selection (reviewed in Ellegren and Galtier 2016). The effect of selection against weakly deleterious mutations on linked neutral variants – Background selection (Charlesworth *et al.* 1993) – can be well approximated by a simple re-scaling of N_e whereas hitchhiking of beneficial or strongly deleterious mutations has more complex effects because there is not a single re-scaling (Barton 1995; Cvijovic et a. 2018). In the case of beneficial mutations, for instance, the interference depends both on the beneficial effect of the sweeping mutation and on selection acting at linked sites (Barton 1995; Weissman and Barton 2012).

Evidence that linked positive selection and not only direct selection on slightly deleterious and beneficial contributed to the relationship between the fraction of mutations affected by selection and N_e has recently been obtained by Castellano *et al.* (2018). Using two *Drosophila melanogaster* genome resequencing datasets, Castellano *et al.* (2018) tested a prediction of the slightly deleterious model first obtained by Kimura (1979) and then extended by Welch *et al.* (2008). Welch *et al.* (2008) showed that if one considers only deleterious mutations, the logarithm of the ratio of nucleotide diversity at non-synonymous and synonymous amino acid changes is linearly related to the logarithm of the effective population size and that the slope of this log-log regression line is equal to the shape parameter of the Distribution of Fitness Effects (DFE), *β*, if the DFE of deleterious mutations is modeled by a Gamma distribution:

$ln(\pi_N/\pi_S) \approx -\beta\ ln(N_e) + constant$ [Eq. 1a]

Or, rewriting this expectation by using π_S as a proxy for N_e:

$ln(\pi_N/\pi_S) \approx -\beta\ ln(\pi_S) + constant'$ [Eq. 1b]

The second equation holds only if variation in π_S only depends on variation in N_e, and does not depend on variation in mutation rates. It should also be pointed out that the DFE considered here only includes deleterious mutations, as estimated for instance by DFE-alpha (Eyre-Walker and Keightley 2009). A direct test of this prediction using among-species comparison can be problematic if mutation rates cannot be controlled for. To circumvent this problem, Castellano *et al.* (2018) used within genome variation in N_e, under the reasonable assumption that variation in mutation rates are negligible compared to variation in N_e across a genome. They found that the slope was significantly more pronounced than expected under a simple scaling of N_e and simulations indicated that linked positive selection, but not background selection, could explain this discrepancy.

In the present paper, we first confirmed the observed pattern on the set of 59 species used in Chen *et al.* (2017). We then used 11 high quality genomic datasets for which an outgroup is available to test whether the results obtained by Castellano *et al.* (2018) hold more generally and, in particular, in species with much smaller effective sizes than *D. melanogaster*, and with different levels of linkage disequilibrium. While we adopted the same general approach than Castellano *et al.* (2018), our analysis differed from theirs in one important respect. In their study, Castellano *et al.* (2018) only characterized the DFE of deleterious mutations. We, instead, used a newly developed approach, *polyDFE* (Tataru *et al.* 2017), that also considers positive mutations, which is expected to improve the estimation of the shape of the DFE of deleterious mutations and to disentangle the direct effects of both positive and negative selection.

Material & Methods

Genomic data and regression of π_N/π_S over π_S

In a first step we analyzed the 59 species used in Chen *et al.* (2017). In later analyses that required unfolded site frequency spectra, we retained 11 species with high quality genomic datasets and with an available outgroup. These eleven species are given in Table 1. They include both animal and plant species with contrasted levels of nucleotide polymorphism and mating systems. We collected Single Nucleotide Polymorphism (SNPs) in all CDS regions and calculated genetic diversity of 4-fold and 0-fold sites as proxies for polymorphism at synonymous (π_S) and non-synonymous sites (π_N). We applied the same SNP sampling strategy as Castellano *et al.* (2018) in order to remove potential dependency between estimates of π_N/π_S and π_S. In brief, we first split all synonymous SNPs into three groups (S1, S2, and S3) using a hypergeometric sampling based on the total number of synonymous sites. To bin genes and reduce the difference in number of SNPs in each bin, we ranked genes according to their Watterson's estimate of nucleotide diversity (θ_{S1}) and grouped these ranked genes into 20 bins each representing approximately 1/20 of the total number of synonymous SNPs. We then used π_{S2} to estimate the π_N/π_S ratio and π_{S3} as an independent estimate of the genetic diversity of each bin.

We calculated the slope of the linear regression (*l*) of the log-transformed value of the π_N/π_S ratio on the log-transformed value of π_S, using the "lm" function in R (R Core Team 2018). In pilot runs on 59 species (population data of Chen *et al.* (2017)), the estimates of *l* showed extensive variation depending on, among other things, the qualities of genome sequencing, read depth, annotation and SNP calling. Thus, we selected 11 species for which a high-quality genome sequence and an outgroup were available. Individuals were selected from the same genetic background, i.e. admixture or population structure were carefully removed. A series of quality controls for *l* calculation were performed as described in the following. The longest transcript for each gene model was kept only if it contained both start and stop codons (putative full length) and no premature stop codons. SNPs within 5 base pairs were masked to avoid false positive calls. A grid of filtering criteria was also implemented on each species based on sequence similarity against Swiss-Prot database (e-value, bit-score, query coverage) and sequencing quality (sites with low read depth or ambiguous variants). We selected the filtering criteria in order to maximize the adjusted R^2 in the log-log regression of π_N/π_S on π_S. By doing so we aimed to reduce the error introduced by annotation and quality difference between model and non-model organisms. Also, to evaluate the variance introduced by random sampling and grouping of SNPs, we performed 1,000-iteration bootstraps to get the bootstrap bias-corrected mean and 95% confidence intervals for *l* calculations.

Estimates of the distributions of fitness effects

The distribution of fitness effects (DFE) for all mutations across the genome was first calculated by considering only deleterious mutations. We first re-used the DFE parameters estimated in 59 animal and plant species in (Chen *et al.* 2017) that assumes that only neutral and slightly deleterious mutations contribute to genetic diversity. In brief, the probability of neutral/deleterious mutations under different selective strength was modeled using a gamma distribution with mean S_d and shape parameter β. Folded site frequency spectra (SFS) were compared between synonymous and nonsynonymous sites and demography (or any departure from equilibrium) was taken into account for by introducing nuisance parameters (Eyre-Walker *et al.* 2006). The possible issues and merits of this approach compared to those based on an explicit (albeit very simplified) demographic model have been discussed previously and the method introduced by Eyre-Walker *et al.* (2006) has proved to be relatively efficient (Eyre-Walker and Keightley 2007; Tataru e*t al.* 2017). The calculations were carried out using an in-house Mathematica script provided in supplementary S2 file of Chen *et al.* (2017).

However, for species with large effective population sizes, like *D. melanogaster*, ignoring the effects of beneficial mutations could distort the DFE to a great extent and lead to a wrong estimate of β. Therefore, we further estimated the DFE under a full model that takes both deleterious and beneficial mutations into account (Tataru *et al.* 2017). The model mixes the gamma distribution of deleterious mutations (shape=β, mean=S_d) with an exponential distribution of beneficial mutations (mean=S_b), in proportions of (1-p_b) and p_b, respectively. The unfolded SFS was calculated for the 11 retained species, for which a closely related outgroup with similar sequencing quality was available to polarize the SFS. The "gamma" DFE (that only considers deleterious mutations) and the full DFE were estimated for each species. In both cases a nuisance parameter was also fitted to account for possible mis-assignment errors in SNP ancestral allele estimation (a step required to obtained the unfolded SFS). Parameters (β, S_b, S_d, and p_b) were estimated using a model averaging procedure where each parameter of interest is estimated as a weighted mean of estimates obtained under the Gamma DFE and full DFE models. The weights given to each estimate reflect the differences in the Akaike Information Criterion (AIC) scores of the Gamma DFE and full DFE models (Posada and Buckley 2004). Calculations were performed using the software *polyDFE* (Tataru *et al.* 2017).

Expectations under different selection models

Independently to possible indirect effects of selective sweeps, [Eq. 1] only considers deleterious mutations, in line with the initial view of the Nearly Neutral Theory where beneficial mutations negligibly contribute to polymorphism (Ohta 1973). Giving more weight to beneficial mutations slightly modified the relationship between the slope of the linear regression, *l,* and the shape parameter, β. For beneficial mutations only, the equivalent of [Eq. 1] is simply (see Appendix):

$$ln(\pi_N/\pi_S) \approx +\beta_b\, ln(N_e) + constant \quad \text{[Eq. 2]}$$

where β_b is the shape of the distribution of beneficial mutations (still assuming a gamma distribution). Thus, the π_N/π_S ratio increases with N_e, so that considering beneficial mutations the global π_N/π_S decreases more slowly than when only deleterious mutations are taken into account. Thus, with beneficial mutations the slope will always be lower than without. For the majority of species beneficial mutations are rare ($\boxed{p_b \ll 1}$) and thus -l is approximately equal to β. For those with a relatively high proportion of beneficial mutations, direct positive selection should result in a flattened slope, i.e. a smaller value of -l than β. As we mostly observed the reverse pattern, -l > β, the observed discrepancy cannot be explained by the direct effect of beneficial mutations.

Trends across the genome and tests for selection

For each of the 20 bins defined above and ranked according to their mean synonymous nucleotide diversity we calculated β, p_b and S_b values and a summary statistic of the site frequency spectrum, Tajima's D (Tajima 1989). Tajima's D tests for an excess of rare over intermediate variants compared to the frequencies expected under the standard coalescent. Demography does affect Tajima's D and can explain the difference among species. However, a negative Tajima's D is also expected under recurrent selective sweeps (Jensen *et al.* 2005; Pavlidis and Alachiotis 2017) and should be more negative in genomic regions more strongly affected by linked positive selection. Background selection can also affect Tajima's D in the same direction but much more weakly (Charlesworth et al. 1995). Independently of the species mean value, we thus expect a strong positive relationship between recombination and Tajima's D in species where linked positive selection is prominent.

Forward simulations under selective sweep scenario

The code developed by Castellano et al (2018) which is based on forward simulations using software SLiM, version 3.2.1 (Haller and Messer 2019) was modified to assess the effect of parameters p_b, S_b, and N_e on –l and Tajima's D. More specifically, a 20-kb non-recombining genomic region was simulated with a mutation rate of 1x10^{-6} to study the behavior of -*l* and Tajima's D under selective sweep scenarios with varying parameters of p_b, S_b, and N$_e$. First, we simulated equal amounts of neutral and deleterious mutations whose fitness effects were drawn from a gamma distribution with a shape parameter 0.4 and a mean s_d of -10. Different percentages of beneficial mutations (p_b= 1%, 0.8%, 0.5%, 0.4%, 0.3%, 0.2%, 0.01%, and 0.005%, 0) were drawn randomly from a distribution with a fixed s_b of 1 to simulate loci experiencing selective sweeps at different frequency and we then calculated -*l* (Fig. 5 of Castellano et al (2018)) and Tajima's D. We also investigated the behavior of -*l* and Tajima's D by varying s_b (1, 0.5, 0.1). Simulation samples were taken after an initial burn in period of 1000 generations and values were averaged across 20 runs.

Results

-l and β are generally similar but the variance is large

One of the most important predictions of the Nearly Neutral Theory is that the proportion of effectively neutral mutations is a function of the effective population size (Kimura and Ohta 1971; Ohta 1972; Ohta 1973; Ohta 1992). In species with large effective population size, selection is efficient and the proportion of effectively neutral mutations is small. Here we used the ratio of genetic diversity at 0-fold over 4-fold degenerate sites (π_N/π_S) in protein coding regions as a measure of the proportion of effectively neutral mutations and examined the linearity between log(π_N/π_S) and log(N$_e$) across the genomes of 59 species used in Chen *et al.* (2017). Although less than half of the species showed a significant regression coefficient (p-value<0.05), the coefficients were negative for 51 of them (*l*<0). The value of *l* varied from -0.424 (*D. melanogaster*) to 0.22 (*Callithrix jacchus*) and the linear relationship between log(π_N/π_S) and log(N$_e$) was statistically significant in 28 species (Table S1). Since balancing selection can lead to both high π_S and π_N/π_S, it can generate an increase in π_N/π_S for high-π_S bins. We thus removed the five bins with the highest diversity and recalculated *l* values for all species. This reduced the *l* values of 36 species and led to negative *l* values in 55 species.

We further examined the DFE for mutations across the genome in the same datasets. A gamma distribution with two parameters, mean (S_d) and shape (β), was used to describe the distribution of deleterious mutations under purifying selection. Importantly, the contribution of beneficial mutations, even those under weak selection that are potentially behaving neutrally, is ignored in this case. Estimates of the shape parameter, β, varied from 0.01 (*C. jacchus*) to 0.347 (*D. melanogaster*) but were only weakly correlated with effective population size (Table S1).

Considering only deleterious mutations and assuming a simple scaling of N_e variation across the genome, the slightly deleterious model predicts that the value of the slope of the linear regression between $\log(\pi_N/\pi_S)$ and $\log(N_e)$, *-l,* is equal to β (Welch *et al.* 2008). The discrepancy between the two might indicate a departure from this model, and Castellano *et al.* (2018) suggested that in *D. melanogaster,* where the observed slope was much larger than β**,** the departure was caused by linked positive selection across the genome. We observed a general consistency between β and *-l* as estimators of effective neutrality (linear coef. = 1.04, intercept=0.007, p-value<2e-16, adjusted R^2=0.35, Fig. 1A). The difference (Δ=-l–β) was small in 40 species dataset and varied from -0.1 to 0.1 (Fig. 1B). In 36 species (61%) *-l* values were larger than β and in 23 species (39%) β was larger than *-l*. However, the variance of Δ was not explained by π_S or N_e as the adjusted R^2 was only 0.06. Removing the five bins with the highest diversity, the correlation between β and *-l* was still significant (coef. 0.89, p-value=2.14e-6). The median value of Δ increased from 0.0085 to 0.045 but there was still no correlation between Δ and N_e.

The effects of quality control and full DFE model

The variance in Δ may come from two sources. First, it can be due to the estimation quality of *-l* and β. Tests have shown that quality control on sequencing and SNP-calling can have a dramatic influence on *-l* calculations and ignoring beneficial mutations in DFE model could also distort the estimates of β

(Tataru *et al.* 2017). Second, the variance in Δ can be caused by departures from the assumptions underlying the simple version of the Nearly Neutral Theory, for instance a larger role of direct or linked positive selection than assumed by the theory.

To assess the relative importance of these two sources we selected 11 species with genomic data of high quality and performed a series of stringent quality controls (see details in M&M) before re-estimating *-l*. This improved the goodness of fit for the log linear regression between π_N/π_S and π_S across the genome and *-l* estimates were significantly different from zero for all 11 species (Table 1, see also details in Table S2). For estimating β, we used closely related species to polarize the SFS and applied both the gamma DFE model and the full DFE model implemented in *polyDFE*, which considers both deleterious and beneficial mutations. Instead of choosing the best DFE model, an average value weighted by the different models' AIC scores was calculated for each parameter (Tataru and Bataillon 2019).

The linear regression model in this case explained a much higher proportion of the variance between *-l* and β (adjusted R^2=0.477) than when we considered the 59 species and used only a gamma DFE. In addition, considering beneficial mutations slightly increases β estimates, making them closer to *–l*. However, the linear coefficient between *-l* and β (1.26) is significantly higher than one and the variation of Δ remained large (-0.026 ~ 0.289) suggesting that some additional factors may lie behind the remaining variation.

The roles of effective population size and positive selection

We then tested if the variation in Δ, where $\Delta = -l - \beta$, could simply reflect differences in effective population size (N_e) among species. Estimates of N_e were obtained by rescaling π_S using estimates of the mutation rate (μ) from the literature. When Δ is regressed against log(N_e), log(N_e) explained up to 49% of the variance in Δ (p-value=0.014). Considering the uncertainty in μ, we also regressed Δ on log(π_S), and obtained similar results (R^2=0.41, p-value=0.019, Fig.2).

Furthermore, we tested whether species with potentially more selective sweeps show higher Δ, as predicted by Castellano *et al.* (2018). An explicit model of selective sweeps is difficult to fit given the uncertainty about beneficial mutations parameters and would require additional information, especially on the recombination map of the different species. Alternatively, we qualitatively reason that, in addition to be more frequent when the effective population is large, the number of selective sweeps should increase with both the proportion (p_b) and the mean strength of beneficial mutations (S_b). Log(S_b) had a significant and positive effect on Δ (p-value=0.0018, Fig. 2) and explained 64.3% of the variance in Δ but the effect of p_b was not significant (p-value=0.29). When considered together, the effects of both log(S_b) and log(π_S) (or Ne) in the joint model explained up to 78% of the variance in Δ (p-value=0.0068 and 0.059, respectively, Table 2). However, no significant effect of p_b could be detected either in the single regression model (p-value=0.29) or joint model with other variables (p-value=0.15).

Trends across the genome and tests for selection

Variation of DFE parameters across bins could also explain the difference between β and $-l$ as the underlying assumptions is that β is constant across bins. We thus calculated β for all 20 bins for the 11 species. Seven species had β values increasing weakly with genetic diversity (p-value<0.05, mean coef.=0.056) while *C. grandiflora* and *H. timareta* had a much faster increase (coef.=0.2 and 0.15, respectively, Table 3). In five species, the maximum β value was still lower than the slope, similar to what was obtained by Castellano *et al.* (2018) in *Drosophila*. However, the maximum β value was larger than the slope in the six remaining species and in five cases the maximum β value was larger than 1 (Table 1). We also compared p_b and S_b values across bins. In *A. thaliana* p_b increased slowly with diversity whereas in *C. grandiflora*, *S. huaylasense*, and *D. melanogaster* p_b decreased significantly (p-value<0.05). In all 11 species, S_b did not show any significant trend across bins. To more formally test for the significance of these variations, we also divided the genomes into five bins (to get enough power per bin) and tested the invariance of the DFE across bins using likelihood ratio tests as implemented in *polyDFE*. For all species, a model with independent DFE parameters for each bin is significantly better than a model with shared parameters across bins (see Table S3).

For all 11 selected species we also calculated Tajima's D (Tajima 1989), thereafter simply called D, in each bin to test for departure from neutrality across the genome. Mean values of D were slightly negative across bins for most species except *S. habrochaites*. For nine of the eleven species, D values increased significantly with genetic diversity (Table 3). Interestingly, we found a negative and strong correlation of Tajima's D with log(S_b) for all 11 species (p-value=0.0086, Pearson's correlation coef. =-0.74) but not with any other DFE parameters. This is in agreement with the expectation that selective sweeps decrease D. We further tested the trends of positive and negative selections by calculating the proportions of deleterious or beneficial mutations over all bins with selective strength <-10 and >10, respectively. However, no significant trends were identified for either kind of direct selections.

We also tested whether alternative measures of the possible occurrence of selective sweeps can also explain the variation in Δ. First we used both the mean Tajima's D and the among-genome correlation between D and π_S (ρ_D) as predictors. More negative D and stronger positive correlation between D and π_S can be viewed as signature of stronger hitchhiking effects. So we predict a negative effect of D and a positive effect of ρ_D. In combination with π_S (or N_e), both D and ρ_D significantly explain variation in Δ (adjusted R^2=0.76, Table 2).

Simulations

Castellano (2018) used forward simulation to assess how *-l* increased under a selective sweep model with varying frequency of adaptive mutations (their Fig. 5). We extended their investigation to assess the effect of selective strength (s_b)

on $-l$ with a fixed β (0.4) and how selective strength (s_b) also affected estimates of Tajima's D. Fig. 3 shows that when s_b increased from 0.1 to 1, $-l$ increased from 0.48 to 0.82 (Δ=0.08 to 0.42). As expected mean Tajima's D decreased as s_b increased but the correlation between D and π_S was only slightly affected (ρ_D, see also Table 4). We also increased N from 100 to 500, and to 1000, and fixed the mean selective strength at either S_b = 10 or S_d = -1000. With these parameters the strength of selection is not affected by N but the number of sweeps increased with N due to the higher input of (beneficial) mutations. In this case Δ increased from 0.079 to 0.75 as N increased and Tajima's D again decreased (Table 4).

Discussion

The aim of the present study was to test quantitatively one of the predictions of the Nearly Neutral Theory of molecular evolution or more precisely the slightly deleterious model, namely that the strength of selection varies with local variations in N_e across the genome depending on the shape of the DFE. We showed that neglecting linked positive selection could lead to a significant quantitative discrepancy between predictions and observations, especially when the effective population size is large. On the other hand, the slightly deleterious model appears as a good approximation when the effective population size is small. Below we first consider possible caveats and discuss the implications of the results for the relative importance of purifying and adaptive selection in shaping the genetic diversity of species.

Caveats: the variation of l and β

In general, estimates of the DFE shape parameter, β, were rather stable compared to estimates of the slope of the regression of $\log(\pi_N/\pi_S)$ over $\log(\pi_S)$, l, with the variance of the former being half that of the latter independently of quality control and whether the SFS was folded or unfolded. High variation in l estimates may explain the fact that a significant correlation between π_N/π_S and π_S could not be observed for all species, particularly those with low genetic diversity (e.g. great apes). Therefore, a stringent quality control for read alignment and SNP calling is necessary even for *D. melanogaster*, where an improvement of the fit in *l* calculation (linear regression adjusted R^2=0.79 to 0.95) leads to a dramatic change in the estimate of Δ (from 0.077 to 0.29). Even if a stringent quality control had been implemented, the goodness of fit for the log linear regression leading to the estimation of *l* would differ significantly from species to species. The fit across the *D. melanogaster* and *A. thaliana* genomes was almost perfect (R^2>0.95) while, at the other extreme, the fit was rather poor in *S. habrochaites* (R^2=0.38). However, even among species for which the fit is almost perfect (R^2>0.95) *l* could vary rather dramatically: *D. melanogaster* had a much larger *l* (0.7) than *A. thaliana* (0.48), *C. rubella* (0.43), and *Z. mays* (teosinte, 0.29), whereas *β* only changed marginally for these species.

On the other hand, we noticed that not all species showed a significant linear relationship between π_N/π_S and N_e or even had positive slopes, especially for those of low diversity (e.g. great apes, Fig 2). Therefore, besides purifying selection *l* is also likely to be affected by additional factors.

A possible source of variance in *β* could be that the single-sided gamma distribution does not describe well the real DFE curves, at least not for all species, particularly when the DFE is not unimodal (Tataru *et al.* 2017). For species like *D. melanogaster*, for instance, there is mounting evidence of adaptive evolution (reviewed in Eyre-Walker 2006). Therefore, it is necessary to consider the possible contribution of beneficial mutations. The full DFE model provided a much better fit than the gamma DFE that considers only deleterious mutations in *D. melanogaster* (log likelihood= -187.3 versus -245.7, respectively). This was also true of some of the outcrossing plants like *Capsella grandiflora*, and *Solanum huaylasense*. In all three species *β* estimates increased when estimated with the Full DFE instead of the Gamma DFE, sometimes significantly (from 0.33 to 0.41 in *D. melanogaster* (Rwanda) and 0.15 to 0.31 in *S. huaylasense*) and at other times only marginally (0.27 to 0.30 in *C. grandiflora*). Taking beneficial mutations into account when fitting the shape of the DFE can partly reduce the discrepancy between *β* estimates and the slope of the regression. However, it is not sufficient as Δ was positive in 10 over the 11 focal species we studied.

Considering positive selection improves the prediction

Based on the prediction of the Nearly Neutral Theory with direct positive selection (Equation 2), the proportion of beneficial mutations is the only factor that could alter the relationship between l and β and should always result in a larger β compared to $-l$. However, this is usually not the case as, on the contrary, values of $-l$ larger than β have generally been reported (Chen *et al.* 2017; James *et al.* 2017; Castellano *et al.* 2018). In this paper we systematically investigated this relationship across the genomes of multiple species. Two thirds of the 59 species and 10 out of the subset of eleven species that were selected for the high quality of their genome, had larger $-l$ than β values. Hence direct positive selection is not the main cause of the discrepancy.

Investigation of DFE parameters changes across bins may help to identify changes in natural selection. Increasing β values over bins could be a signal for stronger positive selection in low diversity regions. Although the maximum β value of some species can be larger than $-l$, β grows slowly for most species and shows hardly any pattern between species. Neither did p_b or S_b. This lack of significant trend in these parameters could simply be due to an increase in variance of their estimates as only one twentieth of the total number of polymorphic sites were used for DFE calculations in each bin. It could also again suggest that direct selection is not the main cause of the discrepancy.

One of the main findings of the present study is that a large proportion of variance in the discrepancy can be explained by the estimated strength of positive selection, which can be regarded as an indication for linked selection, such as selective sweeps or more generally hitchhiking effects. To test for that, we compared changes in Tajima's D and its among-genome correlation coefficients over bins. As expected we observed a negative effect of D and a positive effect of ρ_D on Δ, both suggesting the presence of linked selection, with lower diversity at nearby sites and thus increased discrepancy between $-l$ and β. This is also in agreement with our simulations and those of Castellano et al.

(2018) that illustrate that hitchhiking effects can lower the genetic diversity at nearby neutral or nearly neutral positions. These results can be understood because selective sweep effects cannot simply be captured by a rescaling of N_e. Selective sweeps not only reduce genetic diversity at linked sites but also distort the coalescent genealogy (Fay and Wu 2000; Walsh and Lynch 2018; Campos and Charlesworth 2019), so that we cannot define a single N_e in this context (Weissman and Barton 2012). In particular, the scaling is not expected to be the same for neutral or weakly selected polymorphisms. However, as far as we know, there is no quantitative model predicting the value of the slope as a function of DFE, rates of sweep and recombination rates, and such models still need to be developed.

Conclusions

There are three major conclusions to the present study. First, the Nearly Neutral Theory in its initial form may not explain all aspects of polymorphisms but, almost 50 years after it was first proposed by Tomoko Ohta (Ohta 1973), it still constitutes an excellent starting point for further theoretical developments (Galtier 2016; Walsh and Lynch 2018). Second, considering linked beneficial selection indeed helps to explain more fully polymorphism data, and this is especially true for species with high genetic diversity. This can explain both patterns of synonymous polymorphism (Corbett-Dettig et al. 2015) and how selection reduces non-synonymous polymorphism (Castellano et al. 2018, this study). One could have a progressive increase of the effect of selective sweeps as suggested by Walsh and Lynch (2018, chapter 8) with a shift from genetic drift to genetic draft (Gillespie 1999; 2000; 2001). If so, we could have three domains. For small population sizes, drift would dominate and the nearly neutral theory in its initial form would apply. For intermediate population sizes beneficial mutations would start to play a more important part, and finally for large population sizes, the effect of selective sweeps would dominate and draft would the main explanation of the observed pattern of diversity. Third, our study once more emphasizes the central importance of the DFE in evolutionary genomics and we will likely see further developments in this area.

Acknowledgements: We thank Thomas Bataillon and David Castellano for comments on earlier versions of the manuscript. The project was in part supported by grants from the Swedish Research Council and the Swedish Foundation for Strategic Research to ML.

Table 1 Species and datasets used in the present study

Species	Ref.	Outgroup	Ref.	Mating type	AIC	*l*	β_{full}	β_{gamma}	β_{max}
A. thaliana	Alonso-Blanco *et al.* (2016)	*A. lyrata*	(Novikova *et al.* 2016)	selfing	231.3, 227.3	0.48	0.32	0.32	0.45
A. lyrata	(Novikova *et al.* 2016)	*A. thaliana*	Alonso-Blanco *et al.* (2016)	outcrossing	247.4, 243.4	0.50	0.35	0.34	0.36
C. rubella	(Koenig *et al.* 2018)	*C. grandiflora*	(Agren *et al.* 2014)	selfing	201.4, 200.3	0.43	0.39	0.26	2.86
C. grandiflora	(Agren *et al.* 2014)	*C. rubella*	(Koenig *et al.* 2018)	outcrossing	**321.9,** 327.8	0.52	0.30	0.27	0.36
S. habrochaites	Aflitos *et al.* (2014)	*S. lycopersicon*	Aflitos *et al.* (2014)	selfing	**141.5,** 148.1	0.21	0.23	0.13	3.61
S. huaylasense	Aflitos *et al.* (2014)	*S. lycopersicon*	Aflitos *et al.* (2014)	outcrossing	**87.1,** 121.5	0.54	0.31	0.15	3.89
S. propinquum	Mace *et al.* (2013)	*S. bicolor*	Mace *et al.* (2013)	selfing	163.8, 159.8	0.37	0.26	0.26	0.34
Z. mays (teosinte)	Chia *et al.* (2012)	*T. dactyloides*	Chia *et al.* (2012)	outcrossing	208.1, 204.1	0.29	0.19	0.18	0.45
P. trichocarpa	Evans *et al.* (2014)	*P. nigra*	(Faivre-Rampant *et al.* 2016)	outcrossing	318.9, 319.6	0.42	0.22	0.16	2.21
D. melanogaster	Huang *et al.* (2014)	*D. simulans*	Stanley and Kulathinal (2016)	outcrossing	**422.7,** 535.5	0.70	0.41	0.33	0.51
H. timareta	Martin *et al.* (2013)	*H. melpomene*	Martin *et al.* (2013)	outcrossing	208.2, 204.2	0.44	0.21	0.21	2.78

Note: AIC values were estimated by *polyDFE* for models with and without the effects of beneficial mutations, respectively (bold numbers showed significance < 0.05). So it is with β_{full} and β_{gamma} as well. β_{max} were the maximum value of those estimated by *polyDFE* for each ranked gene bin.

Table 2 Summary table of multiple regression analyses of the effects of π_S S_b, Tajima's D, and ρ_D on Δ, the difference between -*l* and β.

$\Delta \sim \pi_S + \log_{10}(S_b)$	***Coef.***	***SE***	***t value***	***p-value***
Intercept	0.14	0.031	4.69	0.0016**
π_S	7.93	2.96	2.68	0.028*
$\log_{10}(S_b)$	0.015	3.6e-3	4.24	0.0029**
p-value: 0.0008144	Adjusted R^2: 0.7888			
$\Delta \sim \pi_S + D + \rho_D$				
Intercept	-0.031	0.035	-0.87	0.41
Tajima's D	-0.10	0.042	-2.39	0.048*
ρ_D	0.0015	6.05e-4	2.56	0.038*
π_S	15.80	3.39	4.65	0.0040**
p-value: 0.002978	Adjusted R^2: 0.708			

***: $p<0.001$, **: $0.001<p<0.01$, *: $0.01<p<0.05$, ·: $0.05<p<0.1$

Table 3 Changes of summary statistics and DFE parameters across 20 rank gene groups.

	Tajima's D		ρ_{β} ^{a}	ρ_{p_b} ^{a}
	median	ρ_{D}^{a}		
A. thaliana	-0.38	20.10***	0.033***	9.65e-4**
A. lyrata	-0.60	30.13***	0.057*	7.75e-5
C. rubella	-0.28	15.75*	0.039*	8.26e-4
C. grandiflora	-1.06	23.02**	0.20***	-3.53e-3$^{\bullet}$
S. habrochaites	0.22	-5.36	0.11	-7.48e-3
S. huaylasense	-0.17	-8.59**	-0.32	-5.54e-2***
S. propinquum	-0.10	60.04***	0.075***	1.82e-3
Z. mays	-0.52	-0.39	0.055***	2.39e-3
P. trichocarpa	-0.43	79.20***	0.079	-2.80e-3
D. melanogaster	-0.73	7.41**	0.078***	-3.81e-3***
H. timareta	-0.10	6.58**	0.15***	9.87e-4

a: ρ is the slope of the regression of D (β, and p_b, respectively) over genetic diversity across ranked groups of genes.

***: p<0.001, **: 0.001<p<0.01, *: 0.01<p<0.05, $^{\bullet}$: 0.05<p<0.1

Table 4 Results of forward simulations showing the effect of linked positive selection on –l, Δ and summary statistics of the site frequency spectrum for different values of the mean selective value of beneficial mutations, S_b and the population size, N. ρ_D is the correlation between π_S and Tajima's D.

N	S_b	S_d	β	$-l$	Δ	π_S	π_N / π_S	ρ_D	Tajima D
100	20	1000	0.4	-0.485	0.085	0.00136	0.107	9.43E-04	-0.00029
100	50	1000	0.4	-0.664	0.264	0.00122	0.130	1.02E-03	-0.00033
100	100	1000	0.4	-0.822	0.422	0.00101	0.161	8.39E-04	-0.00042
100	10	1000	0.4	-0.479	0.079	0.00153	0.100	1.12E-03	-0.00024
500	10	1000	0.4	-0.491	0.091	0.00580	0.096	1.28E-03	-0.00047
1000	10	1000	0.4	-0.749	0.349	0.00948	0.100	1.20E-03	-0.00058

Figures

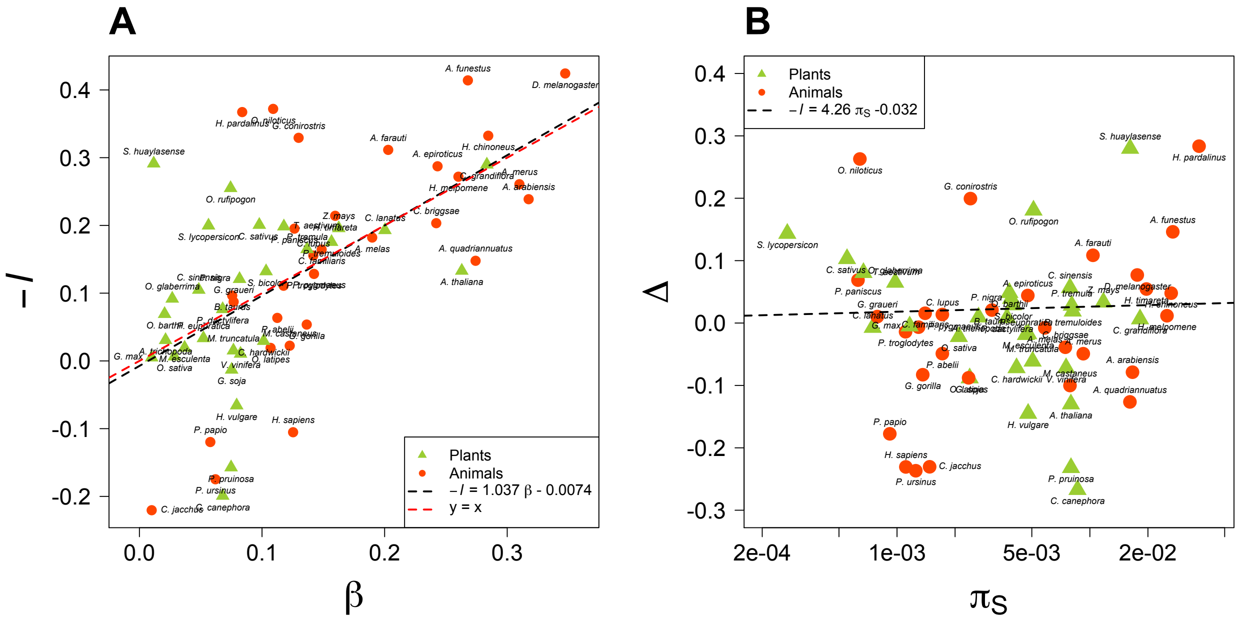

Fig. 1 (A) The correlation between the observed slope of the regression of log(πN/ πS) over πS, -l, and the shape parameter of the DFE, *β,* from the 59 species in Chen *et al.* (2017). (B) The distribution of Δ (=-l-*β*) against genetic diversity at synonymous sites. *β* values were estimated from DFE models with only deleterious mutations considered (the gamma distribution).

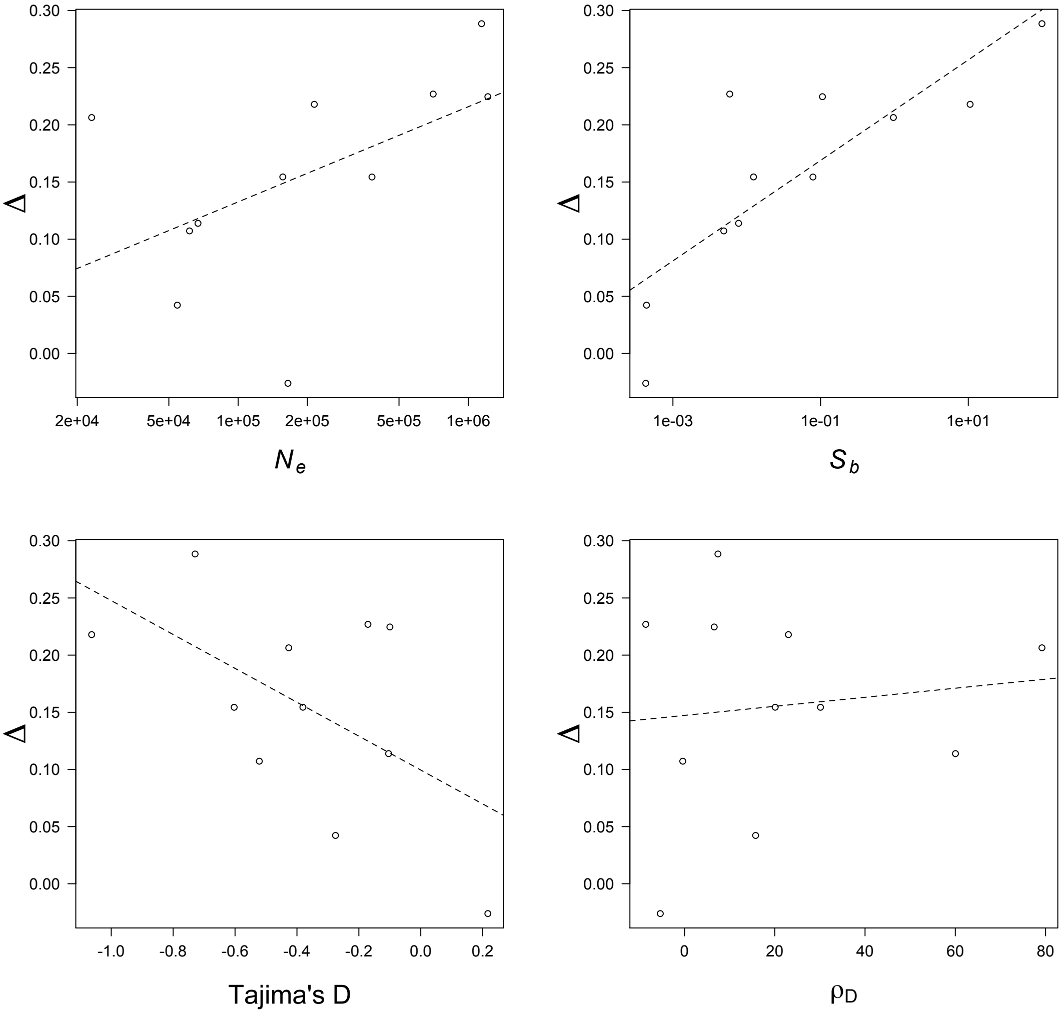

Fig. 2 The relationship between Δ (=-*l*-*β*) and effective population size N_e, selective strength S_b, Tajima's D and the trend of D across bins ρ_D for 11 selected species. Dotted lines showed the linear regression line. *β* and S_b values were estimated from full DFE models with both deleterious and beneficial mutations considered (full DFE model with both gamma and exponential distributions).

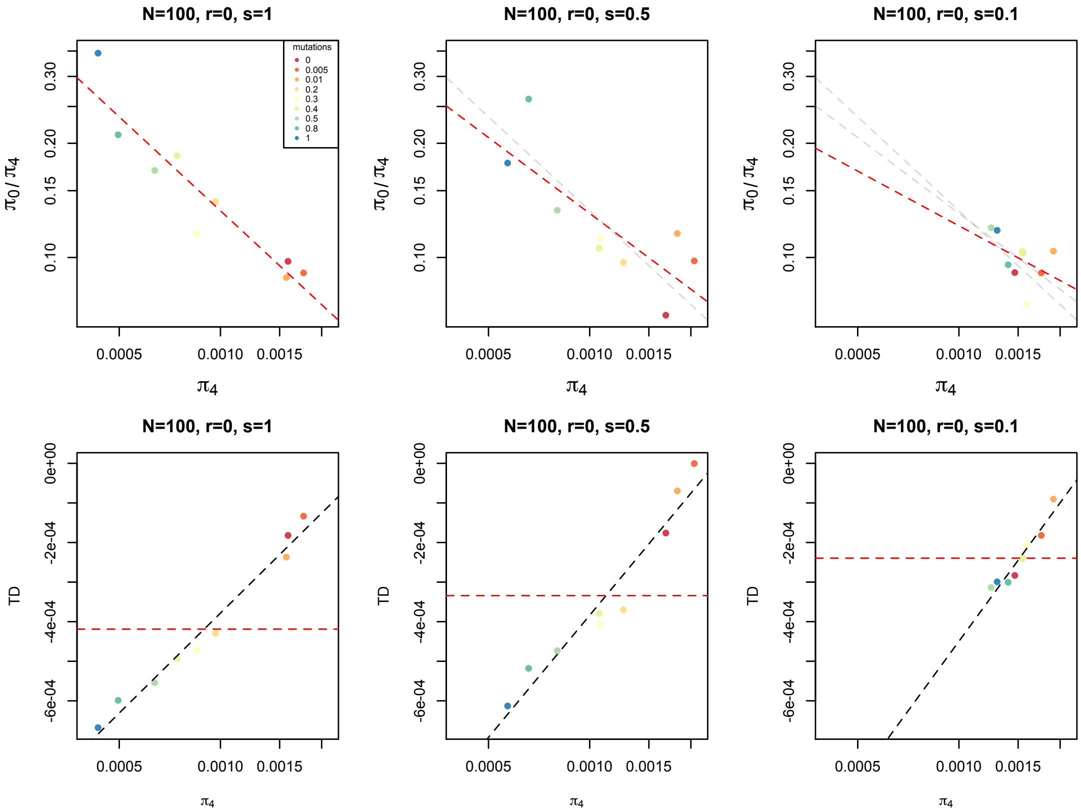

Fig. 3 Effect of linked positive selection on the relationship between log(πN/πS) and log(Ne) and Tajima's D. Upper row: The linear regression coefficient (-*l*) between log(πN/πS) and log(Ne) increases with increasing positive selective strength (from left to right). The red lines are the regression lines. For s=0,5 and s=0.1 the regression lines corresponding to larger s values are indicated with gray lines. Lower row: The red lines for Tajima's D panels indicate the mean values.

Supplementary Information

Supplementary table legends

Table S1. The 59 species used to compare the difference between -*l* and *β* assuming a gamma model for DFE. See Chen et al. (2017) for further details.

Table S2. Details of the 11 species used in the current study to compare the difference between -*l* and β assuming a full model (gamma + exponential) for the DFE.

Table S3. Test for the invariance of DFE parameter estimates across bins by comparing the log-likelihoods of independent estimates for each bin against those of shared estimates.

References

Aflitos, S., E. Schijlen, H. de Jong, D. de Ridder, S. Smit *et al.*, 2014 Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. Plant Journal 80**:** 136-148.

Ågren, J. A., W. Wang, D. Koenig, B. Neuffer, D. Weigel *et al.*, 2014 Mating system shifts and transposable element evolution in the plant genus Capsella. Bmc Genomics 15:602.

Alonso-Blanco, C., J. Andrade, C. Becker, F. Bemm, J. Bergelson *et al.*, 2016 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell 166**:** 481-491.

Barton, N. H., 1995 Linkage and the Limits to Natural-Selection. Genetics 140**:** 821-841.

Campos, J. L., and B. Charlesworth, 2019 The effects on neutral variability of recurrent selective sweeps and background selection. Genetics (in press).

Castellano, D., J. James and A. Eyre-Walker, 2018 Nearly Neutral Evolution Across the Drosophila melanogaster Genome. Molecular Biology and Evolution 35: 2685-2694.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The Effect of Deleterious Mutations on Neutral Molecular Variation. Genetics 134**:** 1289-1303.

Chen, J., S. Glemin and M. Lascoux, 2017 Genetic Diversity and the Efficacy of Purifying Selection across Plant and Animal Species. Molecular Biology and Evolution 34**:** 1417-1428.

Chia, J. M., C. Song, P. J. Bradbury, D. Costich, N. de Leon *et al.*, 2012 Maize HapMap2 identifies extant variation from a genome in flux. Nature Genetics 44**:** 803-807.

Corbett-Detig, R. B., D. L. Hartl and T. B. Sackton, 2015 Natural Selection Constrains Neutral Diversity across A Wide Range of Species. Plos Biology 13(4):e1002112..

Cvijovic, I., B.H. Good and M.M. Desai, 2018 The effect of strong purifying selection on genetic diversity. Genetics 209: 1235-1278.

Ellegren, H., and N. Galtier, 2016 Determinants of genetic diversity. Nature Reviews Genetics 17**:** 422-433.

Evans, L. M., G. T. Slavov, E. Rodgers-Melnick, J. Martin, P. Ranjan *et al.*, 2014 Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. Nature Genetics 46**:** 1089-1096.

Eyre-Walker, A., 2006 The genomic rate of adaptive evolution. Trends in Ecology & Evolution 21**:** 569-575.

Eyre-Walker, A., and P. D. Keightley, 2007 The distribution of fitness effects of new mutations. Nature Reviews Genetics 8**:** 610-618.

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. Molecular Biology and Evolution 26**:** 2097-2108.

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.

Faivre-Rampant, P., G. Zaina, V. Jorge, S. Giacomello, V. Segura *et al.*, 2016 New resources for genetic studies in *Populus nigra*: genome-wide SNP discovery and development of a 12k Infinium array. Molecular Ecology Resources 16: 1023-1036.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.

Galtier, N., 2016 Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. Plos Genetics 12(1): e1005774.

Gillespie, J. H., 1999 The role of population size in molecular evolution. Theoretical Population Biology 55: 145-156.

Gillespie, J. H., 2000 Genetic drift in an infinite population: The pseudohitchhiking model. Genetics 155: 909-919.

Gillespie, J. H., 2001 Is the population size of a species relevant to its evolution? Evolution 55: 2161-2169.

Gillespie, J. H., 2004 *Population genetics : a concise guide*. Johns Hopkins University Press, Baltimore, Md.

Haller, B. C., and P. W. Messer, 2019 SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. Molecular Biology and Evolution 36: 632-637.

Huang, W., A. Massouras, Y. Inoue, J. Peiffer, M. Ramia *et al.*, 2014 Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Research 24: 1193-1208.

James, J., D. Castellano and A. Eyre-Walker, 2017 DNA sequence diversity and the efficiency of natural selection in animal mitochondrial DNA. Heredity 118: 88-95.

Jensen, J. D., B.A. Payseur, W. Stephan, C.F. Aquadro, M. Lynch *et al.*, 2018 The importance of the Neutral Theory in 1968 and 50 years on. Evolution 73: 111-114. .

Jensen, J. D., Y. Kim, V. B. DuMont, C. F. Aquadro and C. D. Bustamante, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170: 1401-1410.

Jensen, J. D., B. A. Payseur, W. Stephan, C. F. Aquadro, M. Lynch *et al.*, 2019 The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. Evolution 73: 111-114.

Kern, A. D., and M. W. Hahn, 2018 The Neutral Theory in Light of Natural Selection. Molecular Biology and Evolution 35: 1366-1371.

Kimura, M., 1979 Model of Effectively Neutral Mutations in Which Selective Constraint Is Incorporated. Proceedings of the National Academy of Sciences of the United States of America 76: 3440-3444.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cabridge, UK: Cambridge Univ. Press.

Kimura, M., and T. Ohta, 1971 Protein Polymorphism as a Phase of Molecular Evolution. Nature 229: 467-469..

Koenig, D., J. Hagmann, R. Li, F. Bemm, T. Slotte *et al.*, 2018 Long-term balancing selection drives evolution of immunity genes in Capsella. eLife 8:e43606.

Kreitman, M., 1996 The neutral theory is dead. Long live the neutral theory. Bioessays 18: 678-683.

Lewontin, R. C., 1974 *The Genetic Basis of Evolutionary Change*. New York: Columbia University Press.

Mace, E. S., S. S. Tai, E. K. Gilding, Y. H. Li, P. J. Prentis *et al.*, 2013 Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nature Communications 4:3320.

Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters *et al.*, 2013 Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Research 23: 1817-1828.

Nei, M., Y. Suzuki and M. Nozawa, 2010 The Neutral Theory of Molecular Evolution in the Genomic Era. Annual Review of Genomics and Human Genetics, Vol 11 11: 265-289.

Novikova, P. Y., N. Hohmann, V. Nizhynska, T. Tsuchimatsu, J. Ali *et al.*, 2016 Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nature Genetics 48: 1077-1082.

Ohta, T., 1972 Population Size and Rate of Evolution. Journal of Molecular Evolution 1: 305-314..

Ohta, T., 1973 Slightly Deleterious Mutant Substitutions in Evolution. Nature 246: 96-98.

Ohta, T., 1992 The Nearly Neutral Theory of Molecular Evolution. Annual Review of Ecology and Systematics 23: 263-286.

Ohta, T., and J. H. Gillespie, 1996 Development of neutral and nearly neutral theories. Theoretical Population Biology 49: 128-142.

Pavlidis, P., and N. Alachiotis, 2017 A survey of methods and tools to detect recent and strong positive selection. Journal of Biological Research-Thessaloniki 24:7.

Posada, D., and T. R. Buckley, 2004 Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. Systematic Biology 53: 793-808.

R Core Team, 2018 R: A language and environment for statistical computing. R Foundation for Statistical Computing, pp. R Foundation for Statistical Computing, Vienna, Austria.

Sawyer, S.A. and D.L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics 132: 1161-1176.

Stanley, C. E., and R. J. Kulathinal, 2016 Genomic signatures of domestication on neurogenetic genes in *Drosophila melanogaster*. Bmc Evolutionary Biology 16:6.

Tajima, F., 1989 Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics 123: 585-595.

Tataru, P., M. Mollion, S. Glemin and T. Bataillon, 2017 Inference of Distribution of Fitness Effects and Proportion of Adaptive Substitutions from Polymorphism Data. Genetics 207: 1103-1119.

Walsh, B., and M. Lynch, 2018 *Evolution and Selection of Quantitative Traits*. Oxford University Press. .

Weissman, D. B., and N. H. Barton, 2012 Limits to the Rate of Adaptive Substitution in Sexual Populations. Plos Genetics 8(6):e1002740.

Welch, J. J., A. Eyre-Walker and D. Waxman, 2008 Divergence and Polymorphism Under the Nearly Neutral Theory of Molecular Evolution. Journal of Molecular Evolution 67: 418-426.

APPENDIX

In a constant population with population size N_e, $\pi_S = 4N_e\mu$ and π_N is given by (Sawyer and Hartl 1992):

$$\pi_N = 2N_e\mu \int_0^1 2x(1-x)H(S,x)dx \quad \text{(A1)}$$

where

$$H(S,x) = \frac{1-e^{-S(1-x)}}{x(1-x)(1-e^{-S})} \qquad \text{(A2)}$$

is the mean time a new semidominant mutation of scaled selection coefficient $S = 4N_es$ spends between x and $x + dx$ (Wright 1938). For constant selection S, by integrating (A1) and dividing by $4N_e\mu$, we have:

$$\frac{\pi_N}{\pi_S} = f(S) = \frac{2}{1-e^{-S}} - \frac{2}{S} \qquad \text{(A3)}$$

(A3) is valid for both positive and negative fitness effect. If we consider only beneficial mutations with a gamma distribution of effects, with mean S_b and shape β_b: $\phi(S_b,\beta,S) = e^{-\frac{S\beta_b}{S_b}} S^{\beta-1} \left(\frac{\beta_b}{S_b}\right)^{\beta_b} / \Gamma(\beta_b)$, we can use the same approach as Welch et al. (2008) to show that:

$$\frac{\pi_N}{\pi_S} = \int_0^\infty f(S)\phi(S_b,\beta_b,S)\,dS$$

$$=\frac{1}{\beta_b-1}\left(\frac{\beta_b}{S_b}\right)^{\beta_b}\left(\xi\left(\beta_b-1,\frac{\beta_b}{S_b}+1\right)+(\beta_b-1)\xi\left(\beta_b,\frac{\beta_b}{S_b}\right)-\xi\left(\beta_b-1,\frac{\beta_b}{S_b}\right)\right) \quad \text{(A4)}$$

where $\xi(x,y)$ is the Hurwith Zeta function. (A4) can be approximated under the realistic assumption that $\frac{\beta_b}{S_b} \ll 1$ and taking Taylor expansion of (A4) in $\frac{\beta_b}{S_b}$ around 0. We thus obtain:

$$\frac{\pi_N}{\pi_S} \approx (2\pi)^{\beta_b}\left(\frac{S_b}{\beta_b}\right)^{\beta_b} \quad \text{(A5)}$$

which leads to equation [eq. 2] in the main text.

species	#chromosom	#genes	slope (l)	R2	p.value
A. trichopoda	8	9002	-0.03079	0.0122	0.4706823
A.thaliana	20	14308	-0.13308	0.4698	0.00382
S.bicolor	7	12382	-0.13228	0.5473	0.00116
M.truncatula	20	7822	-0.015393	0.0081939	0.51873
P.nigra	18	8009	-0.12091	0.19883	0.102593
P.tremula	20	17530	-0.165	0.6679	2.92e-05
P.tremuloides	20	16777	-0.1756	0.6351	4.4e-05
P.euphratica	40	12739	-0.033542	0.03494	0.3758
P.pruinosa	40	15872	0.15714	0.5765	0.000812
V.vinifera	20	10029	-0.010585	-0.002714	0.53938
T. aestivum	5	13135	-0.1985	0.6614	4.00E-04
C.sativus	19	8107	-0.2008	0.37419	0.0346859
C.hardwickii	10	8075	-0.02948	0.005111	0.52121
Z. mays	10	1676	-0.1959	0.292498	0.0379831
G.soja	20	23902	0.01296	0.00965	0.4587
G.max	20	23721	-0.005101	0.004733	0.43566
C.sinensis	4	10983	-0.10496	0.38487	0.0164
O.sativa	20	12416	-0.00658	-0.02686	0.6038
O.rufipogon	11	6305	-0.2551	0.6121	0.001
O.glab	13	8849	-0.09186	0.37368	0.0234
O.barthii	9	6133	-0.06925	0.16368	0.1226006
C.canephora	7	11528	0.1991	0.5222	0.00106
C. lanatus	10	6038	-0.19304	0.18761	0.1244345
M.esculenta	14	12536	-0.01945	0.025044	0.386128
H.vulgare	4	6232	0.06568	0.12179	0.1669793
C. grandiflora	20	12667	-0.2898	0.8196	3.39e-07
P.dactylifera	20	14166	-0.07643	0.240101	0.0538646
S.lycopersicon	5	14665	-0.1998	0.000184	0.6199
S.huaylasense	6	14684	-0.2914	0.000216	0.6211
H.sapiens	20	18191	0.105273	0.046502	0.38661
P.troglodytes	20	16333	-0.12841	0.08777	0.28757
P.paniscus	20	15233	-0.195248	0.12181	0.2393388
G. gorilla	20	12348	-0.053642	0.028298	0.412998
G. graueri	6	13334	-0.087203	0.09082	0.326128
P. abelii	10	15925	-0.06348	0.001645	0.49322
P. pygmaeus	10	15570	-0.11097	0.08732	0.263535
P. Papio	4	13335	0.11973	0.3219	0.0179
P. Ursinus	4	13283	0.1747	0.4082	0.00409
C. Jacchus	10	12859	0.2204	0.5907	0.000123
C. familiaris	20	12670	-0.16431	0.125731	0.2395
C. lupus	8	12665	-0.1555	0.11165	0.199697
B. taurus	18	13714	-0.09646	0.02092	0.37583
O. latipes	20	5478	-0.01919	0.02083	0.4351

O. niloticus	6	4939	-0.372	0.08717	0.1503
G. gorilla	20	2142	-0.3293	0.136386	0.233182
D.melanogaster	20	3686	-0.4243	0.7958	3.94e-06
C. briggsae	10	2497	-0.20332	0.18928	0.1225722
M. castaneus	20	19126	-0.022442	0.006106	0.48217
A. arabiensis	20	6763	-0.2387	0.6917	7.67e-05
A. epiroticus	20	6558	-0.2873	0.5709	0.00118
A. farauti	20	6264	-0.3115	0.4937	0.00448
A. funestus	12	6867	-0.4141	0.7947	5.43e-07
A. melas	12	7148	-0.1821	0.30286	0.0201097
A. merus	20	6665	-0.2608	0.6139	0.000194
A. quadriannuatus	20	6620	-0.1478	0.4223	0.0117
H. melpomene	8	6567	-0.2719	0.7328	4.29e-06
H. chinoneus	8	6437	-0.3324	0.7134	1.83e-05
H. timareta	8	6434	-0.2142	0.6244	0.000243
H. pardalinus	4	6459	-0.3673	0.6618	7.2e-05

S_d	beta
1.85e+22	0.021322045
208.0715692	0.262947282
273522.5492	0.10331407
3520000	0.076671273
7510000	0.081667091
27081.48075	0.13677788
9513.471827	0.156690454
3.00E+08	0.052221393
1740000	0.074766962
2920000	0.082468092
13631.07439	0.117827362
3760000	0.097784926
4390000	0.101323245
18769.08991	0.162600552
5070000	0.075207188
5.35e+48	0.01
1.93e+09	0.048424245
9.12e+13	0.028516702
3190000	0.074351324
8.11e+16	0.026769072
8.94e+23	0.02054373
5640000	0.067941908
357.9045775	0.200255858
1.69e+15	0.036875397
4540000	0.079329761
954.9860899	0.283390544
20400000	0.067902271
43400000	0.056347066
5.97e+53	0.011645271
10002.76568	0.12535229
9517.764368	0.142391471
9478.600653	0.126442144
6536.50279	0.136471801
3920000	0.076851226
176632.7316	0.112477663
49157.40682	0.117581567
66300000	0.057833203
69900000	0.062022659
1.09e+44	0.00999999
58538.91338	0.148629986
123933.2123	0.141824095
1.94e+09	0.075685116
1290000	0.106989872

389319.476	0.109072768
2990000	0.129853019
3974.566141	0.347297266
435.2635717	0.242110286
1890000	0.122589716
3966.393817	0.317476941
3844.612974	0.243166406
27491.72162	0.202916231
35258.79598	0.267911067
29934.62801	0.189863622
2218.056088	0.309945788
10813.37126	0.274193762
15840.97789	0.26019666
9256.968993	0.284517894
1790000	0.159669111
1.49e+12	0.083838116

Species	**filtering***	**#chromosomes**	**slope (l)**	**l_boots(95%)**
A. thaliana	10, 0, 0, 1, 100, 0.1	10	0.477	(0.435, 0.521)
A. lyrata	10, 0, 0, 1, 100, 0.1	10	0.499	(0.585, 0.41)
C. rubella	10, 0, 0,30,100,0.5	10	0.43	(0.387, 0.466)
C. grandiflora	10, 0, 0, 1, 100, 0.5	16	0.521	(0.359, 0.695)
S. habrochaites	1e-20, 200, 0, 1, 100, 0.1	7	0.205	(0.08, 0.319)
S. huaylasense	1e-20, 200, 0, 1, 100, 0.1	4	0.536	(0.423, 0.656)
S. propinquum	10, 0, 0, 1, 10, 0.1	7	0.374	(0.32, 0.426)
Z. mays	10, 0, 0, 1, 100, 1	10	0.292	(0.262, 0.319)
P. trichocarpa	10, 0, 0, 1, 10, 0.1	16	0.421	(0.277, 0.598)
D. melanogaster	10, 0, 0, 1, 10, 0.1	20	0.7	(0.62, 0.768)
H. timareta	10, 0, 0, 1, 100, 0.1	8	0.435	(0.39, 0.476)

*: for filtering we performed following criteria in order: e-value, bit-score, query coverage, quer

different filtering criteria were chose for each species in order to maximize the linearity (R2 colu

p-value	R2	r2_boots(95%)	beta	S_d	p_b	S_b
3.60E-07	0.975	(0.938, 0.995)	0.3225751	-222.003	7.98E-06	0.0123487
4.02E-08	0.88	(0.765, 0.947)	0.3446372	-352.4657	1.95E-05	0.0787558
4.78E-11	0.953	(0.899,0.98)	0.3877494	-280.1572	0.05308512	0.000439121
1.20E-04	0.677	(0.457, 0.839)	0.3030656	-645.0785	0.0117206	10.58446
0.0207338	0.381	(0.0495, 0.69)	0.2311101	-98565.41	0.1430495	0.000428568
3.76E-06	0.79	(0.612, 0.9)	0.309072	-78719.78	0.1444625	0.005878293
1.00E-08	0.904	(0.816, 0.959)	0.260165	-284.4627	2.93E-06	0.007746282
1.30E-10	0.942	(0.893, 0.972)	0.1848003	-2525.62	1.86E-05	0.004883847
1.56E-04	0.679	(0.43, 0.853)	0.2146055	-5353.714	0.03788233	0.9679363
1.26E-11	0.95	(0.909, 0.978)	0.41152	-2175.355	0.00831	99.45153
1.60E-09	0.914	(0.845, 0.961)	0.2103758	-94578.76	0.001163899	0.1059159

y length, num of low quality sites, and percentage of low quality sites
umn) for slope calculation

pi0/pi4	Pi0	Pi4	mutation_	TD	TD0	TD4
0.233257195	0.0010212	0.004378	7	-0.380789691	-0.485021266	-0.164922928
0.183629727	0.0019619	0.010684	7	-0.60281367	-0.742612503	-0.228547355
0.244241748	0.0003722	0.0015239	7	-0.275451105	-0.347128820	-0.145517780
0.2	0.0012	0.006	7	-1.06321752	-1.185242449	-0.706171082
0.203418054	0.0006963	0.003423	5.2	0.216722329	0.194374941	0.251706312
0.175546448	0.00257	0.01464	5.2	-0.17066493	-0.175206441	-0.155516855
0.252857677	0.0006769	0.002677	10	-0.103947788	-0.177934072	-0.000858227
0.331255083	0.002444	0.007378	30	-0.521557494	-0.569127094	-0.375852577
0.220240157	0.000763	0.0034644	37.5	-0.426512977	-0.509090545	-0.155437974
0.090948175	0.0011635	0.012793	2.8	-0.729042402	-1.089469859	-0.273371182
0.109142452	0.00154	0.01411	2.9	-0.099778948	-0.188178832	0.008172753

rhoD	rhoD0	rhoD4
20.09847	13.34179434	21.201422194
30.13018	23.38621	36.52527
15.75497946	15.92928593	3.05971295
23.02078179	9.1258675	46.27447
-5.36317782	-5.25213432	-6.14003074
-8.586076586	-10.405367644	-4.48237127
60.03524	53.42481	63.11604
-0.38718011	-2.28800574	2.64622553
79.19642	66.73075636	85.323726616
7.406745777	-1.309081e+0	12.662693295
6.583048719	3.93145947	6.870488613

Table S3 Test for the invariance of DFE parameter estimates across bins by comparing the log-likelihoods of independent estimates for each bin against those of shared estimates.

Species	**Δ loglk (full DFE)**	Δ Df	p-value	**Δ loglk (gamma DFE)**	Δ Df	p-value
A. thaliana	1361.4	16	0	1005.9	8	0
A. lyrata	1327.6	16	0	695.9	8	3.19e-295
C. rubella	704.8	16	1.45e-290	578.6	8	1.65e-244
C. grandiflora	1018.5	16	0	778.2	8	0
S. habrochaites	204.4	16	5.10e-77	196.4	8	6.54e-80
S. huaylasense	558.5	16	9.27e-228	526.6	8	5.07e-222
S. propinquum	678.4	16	3.30e-279	543.7	8	1.95e-229
Z. mays	721.8	16	7.00e-298	616.3	8	8.84e-261
P. trichocarpa	284.4	16	9.83e-111	307.5	8	1.45e-127
D. melanogaster	169.9	16	1.26e-62	502.7	8	1.07e-211
H. timareta	671.5	16	2.88e-276	543.7	8	2.01e-229