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

1   **From drift to draft: How much do beneficial mutations actually contribute to**

2   **predictions of Ohta's slightly deleterious model of molecular evolution?**

3

4

5   **Jun Chen†,* Sylvain Glémin*,§, Martin Lascoux*,‡**

6

7   †College of Life Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China

8   *Program in Plant Ecology and Evolution, Department of Ecology and Genetics,

9   Evolutionary Biology Centre, Uppsala University, 75236 Uppsala, Sweden

10   § Université de Rennes, CNRS, ECOBIO [Ecosystèmes, Biodiversité, Evolution] -

11   UMR 6553, F-35000 Rennes, France

12

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

1

14   **Abstract**

15   Since its inception in 1973 the slightly deleterious model of molecular evolution,

16   aka the Nearly Neutral Theory of molecular evolution, remains a central model to

17   explain the main patterns of DNA polymorphism in natural populations. This is

18   not to say that the quantitative fit to data is perfect. In a recent study CASTELLANO

19   *et al.* (2018) used polymorphism data from *D. melanogaster* to test whether, as

20   predicted by the Nearly Neutral Theory, the proportion of effectively neutral

21   mutations depends on the effective population size (N_e). They showed that a

22   nearly neutral model simply scaling with N_e variation across the genome could

23   not explain alone the data but that consideration of linked positive selection

24   improves the fit between observations and predictions. In the present article we

25   extended their work in two main directions. First, we confirmed the observed

26   pattern on a set of 59 species, including high quality genomic data from 11

27   animal and plant species with different mating systems and effective population

28   sizes, hence *a priori* different levels of linked selection. Second, for the 11 species

29   with high quality genomic data we also estimated the full Distribution of Fitness

30   Effects (DFE) of mutations, and not solely the DFE of deleterious mutations. Both

31   N_e and beneficial mutations contributed to the relationship between the

32   proportion of effectively neutral mutations and local N_e across the genome. In

33   conclusion, the predictions of the slightly deleterious model of molecular

34   evolution hold well for species with small N_e. But for species with large N_e the fit

35   is improved by incorporating linked positive selection to the model.

36

39

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 the constancy of the molecular clock, 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 considers 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 sweeps). 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.*

73 2015) and is in even stronger contrast with the claim by Kern and Hahn (2018)

74 that "natural selection has played the predominant role in shaping within- and

75 between-species genetic variation" and that "the ubiquity of adaptive variation

76 both within and between species" leads to the rejection of the universality of the

77 Neutral Theory. In a far more nuanced assessment of the Neutral Theory and its

78 contribution, Jensen *et al.* (2019) argued that the effects of linked selection could

79 readily be incorporated in the Nearly Neutral framework. The heart of the

80 dispute, either today or in the early days of the Nearly Neutral Theory, is about

81 the degree to which each category of mutations contributed directly and

82 indirectly to genetic variation within- and between-species.

83

84 A core prediction of the Nearly Neutral Theory is that the fraction of mutations

85 affected by selection depends on the effective population size, N_e (Ohta 1973). N_e

86 is a measure of (the inverse of) genetic drift and corresponds to the size of an

87 ideal, typically Wright-Fisher, population generating the same amount of drift as

88 the observed population. The definition of N_e depends on the drift-related

89 process (ex: inbreeding, variance in allelic frequencies) and the time scale of

90 interest. The coalescent effective population size is relevant for prediction of

91 polymorphism patterns (Sjödin et al, 2005). N_e can vary among species (because

92 of difference in population size, variance in fecundity, reproductive systems…)

93 but also along a genome because linked selection generates variation in the

94 coalescent process if recombination rate and selection targets are not uniformly

95 distributed (reviewed in Ellegren and Galtier 2016). The effect of selection

96 against deleterious mutations on linked neutral variants – background selection

97 (Charlesworth *et al.* 1993) – is often modeled by a simple re-scaling of N_e but

98 except in specific situations effects of linked selection are more complex and

99 there is not a single re-scaling (Barton 1995; Zeng 2013; Comeron 2017; Cvijovic

100 et al. 2018; Torres et al, 2019). In the case of beneficial mutations, for instance,

101 the interference depends both on the beneficial effect of the sweeping mutation

102 and on selection acting at linked sites (Barton 1995; Weissman and Barton 2012).

103

104 Evidence that linked positive selection and not only direct selection on slightly

105 deleterious and beneficial mutations contributed to the relationship between the

106 fraction of mutations affected by selection and N_e has recently been obtained by

107 Castellano *et al.* (2018). Using two *Drosophila melanogaster* genome re-

108 sequencing datasets, Castellano *et al.* (2018) tested a prediction of the slightly

109 deleterious model first obtained by Kimura (1979) and then extended by Welch *et*

110 *al.* (2008). Welch *et al.* (2008) showed that if one considers only deleterious

111 mutations, the logarithm of the ratio of nucleotide diversity at non-synonymous

112 and synonymous amino acid changes is linearly related to the logarithm of the

113 effective population size and that the slope of this log-log regression line is equal

114 to the shape parameter of the Distribution of Fitness Effects (DFE), β, if the DFE

115 of deleterious mutations is modeled by a Gamma distribution:

116

117 $ln(\pi_N/\pi_S) \approx -\beta\, ln(N_e) + constant$ [Eq. 1a]

118

119 where π_N is the nucleotide diversity at non-synonymous sites and π_S is the

120 nucleotide diversity at synonymous sites.

121

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

123

124 $ln(\pi_N/\pi_S) \approx -\beta\, ln(\pi_S) + constant'$ [Eq. 1b]

125

126 The second equation holds only if variation in π_S solely depends on variation in

127 N_e, and that there is no correlation between the mutation rate and N_e. It should

128 also be pointed out that the DFE used here only considers deleterious mutations,

129 as estimated for instance by DFE-alpha (Eyre-Walker and Keightley 2009). A

130 direct test of this prediction using among-species comparison can be problematic

131 if mutation rates cannot be controlled for. To circumvent this problem,

132 Castellano *et al.* (2018) used within genome variation in N_e, under the reasonable

133 assumption that variation in mutation rates are negligible compared to variation

134 in N_e across a genome. They found (see also James et al. 2017) that the slope was

135 significantly steeper than expected under a simple scaling of N_e and simulations

136 indicated that linked positive selection, but not background selection, could

137 explain this discrepancy. The effect of linked selection on the relationship

138 between π_N/π_S and π_S is twofold. First it increases stochasticity in allele

139   frequencies, or, in other words, decreases the local effective population size.

140   Second, linked selection leads to non-equilibrium dynamics. Genetic diversity

141   will recover faster for deleterious than neutral mutations, altering the

142   relationship between π_N/π_S and π_S (Brandvain and Wright, 2016; Do et al. 2015;

143   Gordo and Dionisio 2005; Vigué and Eyre-Walker 2019). More precisely, the

144   more a region is affected by selective sweeps, the lower π_S is and the higher

145   π_N/π_S is compared to the equilibrium expectation: this effect makes the slope

146   steeper compared to the equilibrium expectation.

147

148   In the present paper, we first confirmed the observed pattern on the set of 59

149   species used in Chen *et al.* (2017). We then used 11 high quality genomic

150   datasets for which an outgroup is available to test whether the results obtained

151   by Castellano *et al.* (2018) hold more generally and, in particular, in species with

152   much smaller effective sizes than *D. melanogaster*, and with different levels of

153   linkage disequilibrium. While we adopted the same general approach than

154   Castellano *et al.* (2018), our analysis differed from theirs in one important

155   respect. In their study, Castellano *et al.* (2018) only characterized the DFE of

156   deleterious mutations. We, instead, used a newly developed approach, *polyDFE*

157   (Tataru *et al.* 2017), that also considers positive mutations, which is expected to

158   improve the estimation of the shape of the DFE of deleterious mutations and to

159   disentangle the direct effects of both positive and negative selection.

160

161   **Material & Methods**

162

163   *Genomic data and regression of π_N/π_S over π_S*

164

165   In a first step we re-analyzed the 59 species from Chen *et al.* (2017), which

166   included 34 animals and 28 plant species. We estimated the DFE using folded site

167   frequency spectra with the same method as in Chen *et al.* (2017) and calculated

168   the slope (regression coefficient of $\log(\pi_0/\pi_4)$ over $\log(\pi_4)$ as described in the

169   next paragraph (we used 0-fold degenerated sites for the calculation of genetic

170   diversity at nonsynonymous sites and 4-fold degenerated sites for synonymous

171   sites. The same estimates are used in the rest of the paper). For DFE estimation

using folded SFS the model assumes a gamma distribution for deleterious mutations and takes demography (or sampling or any departure from equilibrium) into account by introducing *n-1* nuisance parameters for an SFS of size *n* (the corresponding code was provided 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. For each of the eleven species, we aligned short reads to the genome using BWA-mem (Li and Durbin 2010) and sorted the alignment using SAMtools. PCR duplicates were removed and INDELs were realigned using GATK toolkit (McKenna *et al.* 2010). HaplotypeCaller was used for individual genotype identification and joint SNP calling was performed across all samples using GenotypeGVCFs. Variant and invariant sites were kept only if genotypes of all individuals were successfully identified (Carson *et al.* 2014). We collected Single Nucleotide Polymorphism (SNPs) in all CDS regions and calculated genetic diversity of 4-fold and 0-fold degenerated sites as proxies for polymorphism at synonymous (π_S) and non-synonymous sites (π_N). Sites were all masked with 'N' and excluded from further computation in the following five cases: heterozygous sites in selfing species, sites with more than two variants, variants at sites within five bases of a flanking INDEL, variants sites with GQ<20, and missing individuals. We applied the same SNP sampling strategy as in James *et al*. (2017) and Castellano *et al.* (2018) in order to remove potential dependency between estimates of π_0/π_4 and π_4. In brief, we first split all synonymous SNPs into three groups (S1, S2, and S3) following a hypergeometric sampling process in R based on the total number of synonymous sites (see formula 3-6 in Castellano *et al* 2019). 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_0/π_4 (we summed π over all genes and scaled by the total length to get π_0 and π_4 for each bin) ratio and π_{S3} as an independent estimate of the genetic diversity of each bin.

7

205   We calculated the slope of the linear regression (l) of the log-transformed value
206   of the π_0/π_4 ratio on the log-transformed value of π_4, using the "lm" function in R
207   (R Core Team 2018). In pilot runs on 59 species (population data of Chen *et al.*
208   (2017)), the estimates of l showed extensive variation depending on, among
209   other things, the qualities of genome sequencing, read depth, annotation and SNP
210   calling. Thus, we selected 11 species for which a high-quality genome sequence
211   and an outgroup were available. Individuals were selected from the same genetic
212   background, i.e. admixture or population structure were carefully removed. At
213   least 20 alleles (i.e. 10 individuals for outcrossing species or 20 for selfing
214   species) were retained from a single ancestral cluster defined in
215   Admixture/Structure analysis in the original publication. For the two *Capsella*
216   species, we performed Admixture analysis for both species separately. A series
217   of quality controls for l calculation were performed as described in the following.
218   The longest transcript for each gene model was kept only if it contained both
219   start and stop codons (putative full length) and no premature stop codons. SNPs
220   flanking five bases of INDEL were masked to avoid false positive calls. A grid of
221   filtering criteria (see details in Table S2) was also implemented on each species
222   based on sequence similarity against Swiss-Prot database (e-value, bit-score,
223   query coverage) and sequencing quality (sites with low read depth or ambiguous
224   variants). We selected the filtering criteria in order to maximize the adjusted R^2
225   in the log-log regression of π_0/π_4 on π_4. By doing so we aimed to reduce the error
226   introduced by annotation and quality difference between model and non-model
227   organisms. Also, to evaluate the variance introduced by random sampling and
228   grouping of SNPs, we performed 1,000-iteration bootstraps to get the bootstrap
229   bias-corrected mean and 95% confidence intervals for l calculations.

230

231   *Estimates of the distributions of fitness effects*

232

233   The distribution of fitness effects (DFE) for all non-synonymous mutations
234   across the genome was first calculated by considering only deleterious
235   mutations. We first re-used the DFE parameters estimated in 59 animal and
236   plant species in Chen *et al.* (2017) that assumes that only neutral and slightly
237   deleterious mutations contribute to genetic diversity. In brief, in this previous

study the DFE 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 by introducing $n-1$ nuisance parameters for an unfolded SFS of size n, following the method proposed by 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 implementing the method of Eyre-Walker *et al.* (2006) 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) using unfolded SFS for 11 species. Briefly, 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. Ancestral state was assigned as the state of the outgroup if the outgroup was monomorphic for one of the two variants, and the derived allele frequency was calculated from this polarization. Otherwise (in the case of missing data, polymorphic site or third allele in the outgroup) the site was masked. The percentage of SNPs that could not be polarized and were masked varied between 0 and 29.3% with a mean of 4.6% and a median value of 0.5% (Table S2).

266   In addition, since polarization errors could remain, the error rate of the ancestral

267   state assignment (ε_{an}) was also taken into account in *PolyDFE*. The "gamma" DFE

268   (that only considers deleterious mutations) and the full DFE were estimated for

269   each species. In both cases a nuisance parameter was also fitted to account for

270   possible mis-assignment errors in SNP ancestral allele estimation (a step

271   required to obtain the unfolded SFS). Note that, although we used outgroups to

272   polarize SFSs, we did not use divergence but only polymorphism to estimate the

273   effect of beneficial mutations. This is at the cost of larger variance in estimates

274   but it avoids the (potentially strong) bias due to ancient variations in N_e that

275   cannot be captured by modeling recent changes in population size (Rousselle et

276   al. 2018). When comparing the estimates of the DFE among several species, the

277   problem arises that the best model is not necessarily the same for all species (the

278   best model can include or not beneficial mutations and include or not

279   polarization errors). Comparisons cannot be fairly done if all species do not

280   share the same model. Alternatively, estimations under an over-parameterized

281   model can lead to large variance and extreme values. To circumvent this problem,

282   we used a model averaging procedure where each parameter of interest (β, S_b, S_d,

283   and p_b) is estimated as a weighted mean of estimates obtained under four models:

284   the Gamma DFE and the full DFE models, including polarization errors or not.

285   The weights given to the estimate from model k is $w_k = e^{-1/2\Delta AIC_k}$ where

286    $\Delta AIC_m = AIC_m - AIC_{min}$ with AIC being the Akaike Information Criterion and

287   AIC_{min} the minimum AIC among the four models (Posada and Buckley 2004). All

288   calculations were performed using the software *polyDFE* and the associated R

289   script (Tataru *et al.* 2017). A goodness-of-fit test was carried out by comparing

290   observed SFS to expected SFS under Gamma DFE or full DFE model for each SNP

291   category, respectively. P-values were calculated under a χ^2 distribution with the

292   number of degrees of freedom equal to n (total number of SNP categories) – 2

293   (synonynmous and non-synonymous sites).

294

295   *Expectations under different selection models*

296

297   Independently of possible indirect effects of selective sweeps, [Eq. 1] only

298   considers deleterious mutations, in line with the initial view of the Nearly

299   Neutral Theory where beneficial mutations negligibly contribute to

300   polymorphism (Ohta 1973). Giving more weight to beneficial mutations slightly

301   modified the relationship between the slope of the linear regression, *l,* and the

302   shape parameter, *β*. For beneficial mutations only, the equivalent of [Eq. 1] is

303   simply (see Appendix):

304

305    $ln(\pi_N/\pi_S) \approx +\beta_b\, ln(N_e) + constant$ [Eq. 2]

306

307   where β_b is the shape of the distribution of beneficial mutations, still assuming a

308   gamma distribution, so β_b would be 1 in the statistical framework we used. Thus,

309   the π_0/π_4 ratio increases with N_e, so that considering beneficial mutations the

310   global π_0/π_4 decreases more slowly than when only deleterious mutations are

311   taken into account. Thus, with beneficial mutations the slope will always be

312   lower than without. For the majority of species beneficial mutations are rare

313   ($\boxed{p_b \ll 1}$) and thus *b* (thereafter we define *b = -l*) is approximately equal to *β*. For

314   those with a relatively high proportion of beneficial mutations, direct positive

315   selection should result in a flattened slope, i.e. a smaller value of *b* than *β*. As we

316   mostly observed the reverse pattern, *b > β*, the observed discrepancy cannot be

317   explained by the direct effect of beneficial mutations.

318

319   *Trends across the genome and tests for selection*

320

321   For each of the 20 bins defined above and ranked according to their mean

322   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 and was calculated from synonymous sites 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 the software SLiM, version 3.2.1 (Haller and Messer 2019) was modified to assess the effect of parameters p_b, S_b, and N on b and Tajima's D. More specifically, a 20-kb genomic region was simulated with a mutation rate of 1×10^{-6} to study the behavior of b and Tajima's D under selective sweep scenarios with varying parameters of p_b, S_b, and N. 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%, 0.005% and 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 b (Fig. 5 of Castellano et al (2018)) and Tajima's D. We also investigated the behavior of b and Tajima's D by varying s_b (1, 0.5, 0.1), N (100, 500, 1000) and the recombination rate (Nr=0, 1e-3, 1e-2). Simulated values were averaged across 50 samples, which were taken every 5N generations after an initial burn-in period of 10N generations.

Results

357 *b and β are generally similar but the variance is large*

358

359 One of the most important predictions of the Nearly Neutral Theory is that the
360 proportion of effectively neutral mutations is a function of the effective
361 population size (Kimura and Ohta 1971; Ohta 1972; Ohta 1973; Ohta 1992). In
362 species with large effective population size, selection is efficient and the
363 proportion of effectively neutral mutations is small. Here we used the ratio of
364 genetic diversity at 0-fold over 4-fold degenerate sites (π_0/π_4) in protein coding
365 regions as a measure of the proportion of effectively neutral mutations and
366 examined the linearity between $\log(\pi_0/\pi_4)$ and $\log(\hat{N}_e)$ across the genomes of
367 59 species used in Chen *et al.* (2017). The slope (linear regression coefficient
368 between $\log(\pi_0/\pi_4)$ and $\log(\hat{N}_e)$) was negative for 51 of the 59 species ($l<0$),
369 although it was significantly different from zero at p=0.05 in less than half of the
370 species (28/59). The value of *l* varied from -0.424 (*D. melanogaster*) to 0.22
371 (*Callithrix jacchus*) (Table S1). Since balancing selection can lead to both high π_4
372 and π_0/π_4, it can generate an increase in π_0/π_4 for high-π_S bins. We thus removed
373 the five bins with the highest diversity and recalculated *l* values for all species.
374 This reduced the *l* values of 36 species and led to negative *l* values in 55 species.

375

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

384

385 Considering only deleterious mutations and assuming a simple scaling of N_e
386 variation across the genome, the slightly deleterious model predicts that the

387　　value of the slope of the linear regression between $\log(\pi_0/\pi_4)$ and $\log(\hat{N}_e)$, b

388　　(recall that $b = -l$), is equal to β (Welch *et al.* 2008). The discrepancy between the

389　　two might indicate a departure from this model, and Castellano *et al.* (2018)

390　　suggested that in *D. melanogaster,* where the observed slope was steeper than

391　　expected, the departure was caused by linked positive selection across the

392　　genome. We observed a general consistency between β and b as estimators of

393　　effective neutrality (linear coef. = 1.04, intercept=0.007, p-value<2e-16, adjusted

394　　R²=0.35, Fig. 1A). The difference ($\Delta=b-\beta$) was small in 40 species and varied

395　　from -0.1 to 0.1 (Fig. 1B). In 36 species (61%) b values were larger than β and in

396　　23 species (39%) β was larger than b. However, the variation in Δ was not

397　　explained by π_S or N_e as the adjusted R² was only 0.06. Removing the five bins

398　　with the highest diversity, the correlation between β and b was still significant

399　　(coef. 0.89, p-value=2.14e-6). The median value of Δ increased from 0.0085 to

400　　0.045 but there was still no correlation between Δ and $\hat{N}_e$.

401

402　　*The effects of quality control and full DFE model*

403

404　　The variation in Δ may come from two sources. First, it can be due to the

405　　estimation quality of b and β. Tests have shown that quality control on

406　　sequencing and SNP-calling can have a dramatic influence on b calculations and

407　　ignoring beneficial mutations in DFE model could also distort the estimates of β

408　　(Tataru *et al.* 2017). Second, the variation in Δ can be caused by departures from

409　　the assumptions underlying the simple version of the Nearly Neutral Theory, for

410　　instance, a larger role of direct or linked positive selection than assumed by the

411　　theory.

412

413　　To assess the relative importance of these two sources we selected 11 species

414　　with genomic data of high quality and performed a series of stringent quality

415　　controls (see details in M&M) before re-estimating b. This improved the

416　　goodness of fit for the log linear regression between π_0/π_4 and π_4 across the

417　　genome and b estimates were significantly different from zero for all 11 species

418　　(Table 1 and Fig. 2, see also details in Table S2 and Fig. S1-S3 for correlation

419   between π_0/π_4 and π_4 and between π_0 and π_4 in linear and log scales). For

420   estimating β, we used closely related species to polarize the SFS and applied both

421   the gamma DFE model and the full DFE model implemented in *polyDFE*, which

422   considers both deleterious and beneficial mutations. Instead of choosing the best

423   DFE model, an average value weighted by the different models' AIC scores was

424   calculated for each parameter (Tataru and Bataillon 2019).

425

426   In this case we observed a better correlation between b and β (Pearson's

427   correlation rho = 0.727, p-value=0.011) than when we considered the 59 species

428   and used only a gamma DFE. In addition, considering beneficial mutations

429   slightly increases β estimates, making them closer to b. However, the linear

430   coefficient between b and β (1.26) is significantly higher than one and the

431   variation of Δ remains large (-0.026 $\sim$ 0.289) suggesting that some additional

432   factors may lie behind the remaining variation.

433

434   *The roles of effective population size and positive selection*

435

436   We then tested if the variation in Δ, where $\Delta=b-\beta$, could simply reflect

437   differences in effective population size (N_e) among species. Estimates of N_e were

438   obtained by rescaling π_S using estimates of the mutation rate (μ) from the

439   literature (see Table S3 for the sources of the μ estimates). When Δ is regressed

440   against $\log(\hat{N}_e)$, $\log(\hat{N}_e)$ explained up to 49% of the variance in Δ (p-

441   value=0.014). Considering the uncertainty in μ, we also regressed Δ on $\log(\pi_S)$,

442   and obtained similar results (R^2=0.41, p-value=0.019, Fig. 3).

443

444   Furthermore, we tested whether species with potentially more selective sweeps

445   show higher Δ, as predicted by Castellano *et al.* (2018). An explicit model of

446   selective sweeps is difficult to fit given the uncertainty about beneficial

447   mutations parameters and would require additional information, especially on

448   the recombination map of the different species. Alternatively, we qualitatively

449   reasoned that, in addition to be more frequent when the effective population is

450   large, the number of selective sweeps should increase with both the proportion

451 (p_b) and the mean strength of beneficial mutations (S_b). Log(S_b) had a significant

452 and positive effect on Δ (p-value=0.0018, Fig. 3) and explained 64.3% of the

453 variance in Δ but the effect of p_b was not significant (p-value=0.29). When

454 considered together, the effects of both log(S_b) and log($π_4$) (or $\hat{N}_e$) in the joint

455 model explained up to 78% of the variance in Δ (p-value=0.0068 and 0.059,

456 respectively, Table 2). However, no significant effect of p_b could be detected

457 either in the single regression model (p-value=0.29) or joint model with other

458 variables (p-value=0.15). The rate of adaptive evolution relative to the neutral

459 mutation rate, ω_a (Galtier 2016) combines the proportion (p_b) and the mean

460 strength of beneficial mutations (S_b) according to $\omega_a = p_S \times S_b / (1 - \exp(-S_b))$.

461 However, as for p_b the effect of ω_a on Δ was not significant (p-value=0.17)

462 although the relationship is positive as expected.

463

464 We also conducted goodness-of-fit tests for both Gamma DFE and full DFE

465 models by comparing the difference between the observed SFS to that simulated

466 under both DFE models, respectively (Fig S4, Table S4). For all 11 species, the

467 expected SFS under the full DFE model showed a better or as good fit to the data

468 as the one under the Gamma DFE model (higher or equal *p*-values), especially for

469 species like *C. grandiflora*, *D. melanogaster*, *S. habrochaites*, *S. huaylasense* and *P.*

470 *trichocarpa* where either p_b or S_b could not be ignored. Overall, the full DFE

471 model predicted the observed SFS with less than 5% error (Figure S4).

472

473 *Trends across the genome and tests for selection*

474

475 Variation of DFE parameters across bins could also explain the difference

476 between *β* and *b* since the underlying assumption is that *β* is constant across bins.

477 We thus calculated *β* for all 20 bins for the 11 species. Seven species had *β* values

478 increasing weakly with genetic diversity (p-value<0.05, mean regression

479 coefficient 0.056) while *C. grandiflora* and *H. timareta* had a much faster increase

480 (regression coefficient =0.2 and 0.15, respectively, Table 3). In five species, the

481 slope was steeper than the maximum *β* value, similar to what was obtained by

482 Castellano *et al.* (2018) in *Drosophila*. However, the slope was shallower than the

483 maximum β value in the six remaining species and in five of them the maximum β

484 value was larger than 1 (Table 1). We also compared p_b and S_b values across bins.

485 In *A. thaliana* p_b increased slowly with diversity whereas in *C. grandiflora*, *S.*

486 *huaylasense*, and *D. melanogaster* p_b decreased significantly (p-value<0.05). In all

487 11 species, S_b did not show any significant trend across bins. To more formally

488 test for the significance of these variations, we also divided the genomes into five

489 bins (to get enough power per bin) and tested the invariance of the DFE across

490 bins using likelihood ratio tests as implemented in *polyDFE*. For all species, a

491 model with independent DFE parameters for each bin is significantly better than

492 a model with shared parameters across bins (see Table S5).

493

494 For all 11 selected species we also calculated Tajima's D (Tajima 1989),

495 thereafter simply called D, in each bin to test for departure from neutrality

496 across the genome. Mean values of D were slightly negative across bins for most

497 species except *S. habrochaites*. For nine of the eleven species, D values increased

498 significantly with genetic diversity (Table 3). Interestingly, we found a negative

499 and strong correlation of Tajima's D with log(S_b) for all 11 species (Pearson's

500 correlation rho. =-0.74, p-value=0.0086) but not with any other DFE parameters.

501 This is in agreement with the expectation that selective sweeps decrease D.

502 Background selection could also decrease D albeit to a lower extent. We further

503 tested the trends of positive and negative selection by calculating the

504 proportions of deleterious or beneficial mutations over all bins with selective

505 strength <-10 and >10, respectively. However, no significant trends were

506 identified for either type of direct selection.

507

508 We also tested whether alternative measures of the possible occurrence of

509 selective sweeps could explain a larger part of the variation in Δ. We used both

510 the mean Tajima's D and the among-genome regression coefficient of the

511 relationship between D and π_S (ρ_D) as predictors. More negative D and stronger

512 positive regression coefficient between D and π_S can be viewed as signature of

513 stronger hitchhiking effects. So we would expect to see a negative effect of D and

514 a positive effect of ρ_D on the variation in Δ. In combination with π_4 (or $\hat{N}_e$), both

515 D and ρ_D indeed explained a significant part of the variation in Δ (adjusted
516 R^2=0.76, Table 2).

517

518 *Simulations*

519

520 Castellano et al. (2018) used forward simulations to assess the extent to which

521 selective sweeps made the slope the relationship between $\log(\pi_0/\pi_4)$ and $\log(\hat{N}_e)$

522 steeper and thereby could explain the discrepancy between the slope and the

523 shape parameter of the DFE, β. They tested varying proportions of adaptive

524 mutations (their Fig. 5). We extended their investigation to test the effect of

525 selective strength (s_b) on b with a fixed β (0.4) and how selective strength (s_b)

526 also affected estimates of Tajima's D. Without recombination (Nr=0), Fig. 4

527 shows that when s_b increased from 0.1 to 1, b increased from 0.46 to 0.72

528 (Δ=0.06 to 0.32). As expected mean Tajima's D decreased from -0.36 to -0.77 as

529 s_b increased and ρ_D between D and π_4 increased (see also Table 4). We also

530 increased N from 100 to 500, and to 1000, and fixed the mean selective strength

531 at either S_b = 10 or S_d = -1000. With these parameters, the strength of selection

532 was not affected by N but the number of sweeps increased with N due to the

533 higher input of (beneficial) mutations. In this case Δ increased from 0.06 to 0.41

534 as N increased and Tajima's D again decreased (Table 4 and Fig. 5). With

535 recombination (Nr=1e-3 and Nr=1e-2), we noticed similar trends of b, D, and ρ_D

536 when s_b or N are large enough to recover the significance of the linearity between

537 $\log(\pi_0/\pi_4)$ and $\log(\pi_4)$ (Fig. S5 and S6).

538

539 **Discussion**

540

541 The aim of the present study was to test quantitatively one of the predictions of

542 the Nearly Neutral Theory of molecular evolution or, more precisely, the slightly

543 deleterious model. More specifically, we used full genome datasets to test

544 whether the proportion of effectively neutral mutations varies with local

545 variation in N_e across the genome and decreases linearly with increasing N_e and

546 whether the slope is equal to the shape parameter of the DFE. The negative log

547   linear relationship between π_0/π_4 and N_e observed in previous studies

548   (Gossmann et al. 2011; Murray et al. 2017; Castellano et al. 2018; Vigué and

549   Eyre-Walker 2019) was also observed in the present study, although the slope

550   was not always significantly negative and, when negative, could differ

551   significantly from the shape parameter of the DFE and be much steeper. The

552   latter was especially true in species with large effective population size and the

553   difference was correlated to the estimated mean strength of selection acting on

554   beneficial mutations. In the case of species with large effective population size

555   neglecting linked positive selection could therefore lead to a significant

556   quantitative discrepancy between predictions and observations. On the other

557   hand, the slightly deleterious model appears as a good approximation when the

558   effective population size is small. Below we first consider possible caveats and

559   discuss the implications of the results for the relative importance of purifying

560   and adaptive selection in shaping the genetic diversity of species.

561

562   The discrepancy between the slope of the log linear relationship between π_N/π_S

563   and N_e and β could simply be due to difficulties in estimating them precisely. In

564   general, estimates of the DFE shape parameter, β, were rather stable compared

565   to estimates of the slope of the regression of $\log(\pi_0/\pi_4)$ over $\log(\pi_4)$, b, with the

566   variance of the former being half that of the latter independently of quality

567   control and whether the SFS was folded or unfolded. High variation in b

568   estimates may explain the fact that a significant correlation between π_0/π_4 and

569   π_4 could not be observed for all species, particularly those with low genetic

570   diversity (e.g. great apes). Therefore, a stringent quality control for read

571   alignment and SNP calling is necessary, even for *D. melanogaster*, where an

572   improvement of the fit in *l* calculation (linear regression adjusted R^2=0.79 to 0.95)

573   leads to a dramatic change in the estimate of Δ (from 0.077 to 0.29). Even if a

574   stringent quality control had been implemented, the goodness of fit for the log

575   linear regression leading to the estimation of b would differ significantly from

576   species to species. The fit across the *D. melanogaster* and *A. thaliana* genomes

577   was almost perfect (R^2>0.95) while, at the other extreme, the fit was rather poor

578   in *S. habrochaites* (R^2=0.38). However, even among species for which the fit is

579   almost perfect (R^2>0.95) b could vary rather dramatically: *D. melanogaster* had a

19

580   much larger l (0.7) than *A. thaliana* (0.48), *C. rubella* (0.43), and *Z. mays* (teosinte,

581   0.29), whereas β only changed marginally for these species. Not all species

582   though showed a significant negative linear relationship between π_0/π_4 and $\hat{N}_e$

583   and some even had positive slopes, especially for those of low diversity (e.g.

584   great apes, Fig 2). Therefore, besides purifying selection the slope is also likely to

585   be affected by additional factors. Factors that affect the likelihood to observe a

586   negative relationship between π_0/π_4 and $\hat{N}_e$ and its relationship with the DFE

587   parameters were thoroughly discussed by Castellano et al. (2018). Below we

588   highlight those that seem particularly relevant when considering a group of

589   species with contrasted levels of diversity as was done here. These factors are

590   the variation in N_e estimates along the genome, which itself reflects the joint

591   distribution along the genome of recombination rate and density of selected sites,

592   the DFE, and the variation along the genome of the rate of adaptive evolution

593   (Castellano et al. 2018).

594

595   Lack of joint variation in recombination rate and selected sites seems to be an

596   unlikely cause for an absence of negative relationship between π_0/π_4 and N_e as

597   such a relationship is observed in selfing species where this joint variation is

598   expected to the more limited than in outcrossing ones. A possible source of

599   variance in β could be that the single-sided gamma distribution does not describe

600   well the real DFE curves, at least not for all species, particularly when the DFE is

601   not unimodal (Tataru *et al.* 2017). For species like *D. melanogaster*, for instance,

602   there is mounting evidence of adaptive evolution (reviewed in Eyre-Walker 2006,

603   Sella et al. 2009). Therefore, it is necessary to consider the possible contribution

604   of beneficial mutations. The full DFE model provided a much better fit than the

605   gamma DFE that considers only deleterious mutations in *D. melanogaster* (log

606   likelihood= -187.3 versus -245.7, respectively). This was also true of some of the

607   outcrossing plants like *Capsella grandiflora*, and *Solanum huaylasense*. In all three

608   species β estimates increased when estimated with the Full DFE instead of the

609   Gamma DFE, sometimes significantly (from 0.33 to 0.41 in *D. melanogaster*

610   (Rwanda) and 0.15 to 0.31 in *S. huaylasense*) and at other times only marginally

611   (0.27 to 0.30 in *C. grandiflora*). Taking beneficial mutations into account when

612   fitting the shape of the DFE can partly reduce the discrepancy between β

613   estimates and the slope of the regression. However, it is not sufficient as Δ was

614   positive in 10 over the 11 focal species we studied.

615

616   Based on the prediction of the Nearly Neutral Theory with direct positive

617   selection (Equation 2), the proportion of beneficial mutations is the only factor

618   that could alter the relationship between b and β and should always result in a

619   larger β compared to b. However, this is usually not the case as, on the contrary,

620   values of b larger than β have generally been reported (Chen *et al.* 2017; James *et*

621   *al.* 2017; Castellano *et al.* 2018). In this paper we systematically investigated this

622   relationship across the genomes of multiple species. Two thirds of the 59 species

623   and 10 out of the subset of eleven species that were selected for the high quality

624   of their genome, had larger b than β values. Hence direct positive selection is not

625   the main cause of the discrepancy.

626

627   Investigation of DFE parameter changes across bins may help to identify changes

628   in natural selection. Increasing β values over bins could be a signal for stronger

629   positive selection in low diversity regions. Although the maximum β value of

630   some species can be larger than b, β grows slowly for most species and shows

631   hardly any pattern between species. Neither did p_b or S_b. This lack of significant

632   trend in these parameters could simply be due to an increase in variance of their

633   estimates as only one twentieth of the total number of polymorphic sites were

634   used for DFE calculations in each bin. It could also again suggest that direct

635   selection is not the main cause of the discrepancy.

636

637   One of the main findings of the present study is that a large proportion of

638   variance in the discrepancy can be explained by the estimated strength of

639   positive selection, which can be regarded as an indication for linked selection,

640   such as selective sweeps or more generally hitchhiking effects. To test for that,

641   we compared changes in Tajima's D and its among-genome correlation

642   coefficients over bins. As expected we observed a negative effect of D and a

643   positive effect of ρ_D on Δ, both suggesting the presence of linked selection, with

644   lower diversity at nearby sites and thus increased discrepancy between b and β.

645 This is also in agreement with our simulations and those of Castellano et al.

646 (2018) that illustrate that hitchhiking effects can lower the genetic diversity at

647 nearby neutral or nearly neutral positions. These results can be understood

648 because selective sweep effects cannot simply be captured by a rescaling of N_e.

649 Selective sweeps not only reduce genetic diversity at linked sites but also distort

650 the coalescent genealogy (Fay and Wu 2000; Walsh and Lynch 2018; Campos

651 Parada and Charlesworth 2019), so that we cannot define a single N_e in this

652 context (Weissman and Barton 2012). In particular, the scaling is not expected to

653 be the same for neutral or weakly selected polymorphisms. However, as far as

654 we know, there is no quantitative model predicting the value of the slope as a

655 function of DFE, rates of sweep and recombination rates, and such models still

656 need to be developed.

657

658 **Conclusions**

659

660 There are three major conclusions to the present study. First, the Nearly Neutral

661 Theory in its initial form may not explain all aspects of polymorphisms but,

662 almost 50 years after it was first proposed by Tomoko Ohta (Ohta 1973), it still

663 constitutes an excellent starting point for further theoretical developments

664 (Galtier 2016; Walsh and Lynch 2018). Second, considering linked beneficial

665 selection indeed helps to explain more fully polymorphism data, and this is

666 especially true for species with high genetic diversity. This can explain both

667 patterns of synonymous polymorphism (Corbett-Detig et al. 2015) and how

668 selection reduces non-synonymous polymorphism (Castellano et al. 2018, this

669 study). One could have a progressive increase of the effect of selective sweeps as

670 suggested by Walsh and Lynch (2018, chapter 8) with a shift from genetic drift to

671 genetic draft (Gillespie 1999; 2000; 2001). If so, we could have three domains.

672 For small population sizes, drift would dominate and the nearly neutral theory in

673 its initial form would apply. For intermediate population sizes beneficial

674 mutations would start to play a more important part, and finally for large

675 population sizes, the effect of selective sweeps would dominate and draft would

676 be the main explanation of the observed pattern of diversity. Third, our study

677 once more emphasizes the central importance of the DFE in evolutionary

678 genomics and we will likely see further developments in this area.

679

684

685 **Data availability:** All calculation files can be found in Supplementary file 1. The

686 vcf files used in the present study are available on request.

687

688

689   **References**

690   Aflitos, S., E. Schijlen, H. de Jong, D. de Ridder, S. Smit *et al.*, 2014 Exploring
691   genetic variation in the tomato (Solanum section Lycopersicon) clade by
692   whole-genome sequencing. Plant Journal 80**:** 136-148.
693   Agren, J. A., W. Wang, D. Koenig, B. Neuffer, D. Weigel *et al.*, 2014 Mating system
694   shifts and transposable element evolution in the plant genus Capsella.
695   Bmc Genomics 15.
696   Alonso-Blanco, C., J. Andrade, C. Becker, F. Bemm, J. Bergelson *et al.*, 2016 1,135
697   Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis
698   thaliana. Cell 166**:** 481-491.
699   Barton, N. H., 1995 Linkage and the Limits to Natural-Selection. Genetics 140**:**
700   821-841.
701   Brandvain, Y., & Wright, S. I. 2016 The limits of natural selection in a
702   nonequilibrium world. Trends in Genetics 32:201–210.
703   Campos Parada, J. L., and B. Charlesworth, 2019 The effects on neutral variability
704   of recurrent selective sweeps and background selection. bioRxiv**:** 358309.
705   Carson, A. R., E. N. Smith, H. Matsui, S. K. Braekkan, K. Jepsen *et al.*, 2014 Effective
706   filtering strategies to improve data quality from population-based whole
707   exome sequencing studies. Bmc Bioinformatics 15.
708   Castellano, D., J. James and A. Eyre-Walker, 2018 Nearly Neutral Evolution Across
709   the Drosophila melanogaster Genome. Molecular Biology and Evolution**:**
710   msy164-msy164.
711   Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The Effect of
712   Deleterious Mutations on Neutral Molecular Variation. Genetics 134**:**
713   1289-1303.
714   Chen, J., S. Glemin and M. Lascoux, 2017 Genetic Diversity and the Efficacy of
715   Purifying Selection across Plant and Animal Species. Molecular Biology
716   and Evolution 34**:** 1417-1428.
717   Chia, J. M., C. Song, P. J. Bradbury, D. Costich, N. de Leon *et al.*, 2012 Maize
718   HapMap2 identifies extant variation from a genome in flux. Nature
719   Genetics 44**:** 803-U238.
720   Comeron, J.M. 2017 Background selection as null hypothesis in population
721   genomics: insights and challenges from Drosophila studies. Phil. Trans. R.
722   Soc. B 372: 20160471.
723   Coop, G. 2016 Does linked selection explain the narrow range of genetic diversity
724   across species? bioRxiv doi: https://doi.org/10.1101/042598.
725   Corbett-Detig, R. B., D. L. Hartl and T. B. Sackton, 2015 Natural Selection
726   Constrains Neutral Diversity across A Wide Range of Species. Plos Biology
727   13.
728   Cvijovic, I., B.H. Good and M.M. Desai, 2018 The effect of strong purifying
729   selection on genetic diversity. Genetics 209: 1235-1278.
730   Do, R., D. Balick, H. Li, I. Adzhubei, S. Sunyaev and D. Reich 2015 No evidence that
731   selection has been less effective at removing deleterious mutations in
732   Europeans than in Africans. Nature genetics 47: 126
733   Ellegren, H., and N. Galtier, 2016 Determinants of genetic diversity. Nature
734   Reviews Genetics 17**:** 422-433.

735 Evans, L. M., G. T. Slavov, E. Rodgers-Melnick, J. Martin, P. Ranjan *et al.*, 2014
736   Population genomics of Populus trichocarpa identifies signatures of
737   selection and adaptive trait associations. Nature Genetics 46**:** 1089-1096.
738 Eyre-Walker, A., 2006 The genomic rate of adaptive evolution. Trends in Ecology
739   & Evolution 21**:** 569-575.
740 Eyre-Walker, A., and P. D. Keightley, 2007 The distribution of fitness effects of
741   new mutations. Nature Reviews Genetics 8**:** 610-618.
742 Eyre-Walker, A., and P. D. Keightley, 2009 Estimating the Rate of Adaptive
743   Molecular Evolution in the Presence of Slightly Deleterious Mutations and
744   Population Size Change. Molecular Biology and Evolution 26**:** 2097-2108.
745 Faivre-Rampant, P., G. Zaina, V. Jorge, S. Giacomello, V. Segura *et al.*, 2016 New
746   resources for genetic studies in Populus nigra: genome-wide SNP
747   discovery and development of a 12k Infinium array. Molecular Ecology
748   Resources 16**:** 1023-1036.
749 Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection.
750   Genetics 155**:** 1405-1413.
751 Galtier, N., 2016 Adaptive Protein Evolution in Animals and the Effective
752   Population Size Hypothesis. Plos Genetics 12.
753 Gillespie, J. H., 1999 The role of population size in molecular evolution.
754   Theoretical Population Biology 55**:** 145-156.
755 Gillespie, J. H., 2000 Genetic drift in an infinite population: The
756   pseudohitchhiking model. Genetics 155**:** 909-919.
757 Gillespie, J. H., 2001 Is the population size of a species relevant to its evolution?
758   Evolution 55**:** 2161-2169.
759 Gillespie, J. H., 2004 *Population genetics : a concise guide*. Johns Hopkins
760   University Press, Baltimore, Md.
761 Gordo, I., Dionisio, F. 2005 Nonequilibrium model for estimating parameters of
762   deleterious mutations. *Physical Review. E, Statistical, Nonlinear, and Soft*
763   *Matter Physics* 71: 031907.
764 Gossmann, T. I., Woolfit, M., Eyre-Walker, A. 2011 Quantifying the variation in the
765   effective population size within a genome. *Genetics* 189:1389-1402.
766 Haller, B. C., and P. W. Messer, 2019 SLiM 3: Forward Genetic Simulations
767   Beyond the Wright-Fisher Model. Molecular Biology and Evolution 36**:**
768   632-637.
769 Huang, W., A. Massouras, Y. Inoue, J. Peiffer, M. Ramia *et al.*, 2014 Natural
770   variation in genome architecture among 205 Drosophila melanogaster
771   Genetic Reference Panel lines. Genome Research 24**:** 1193-1208.
772 James, J., D. Castellano and A. Eyre-Walker, 2017 DNA sequence diversity and the
773   efficiency of natural selection in animal mitochondrial DNA. Heredity 118**:**
774   88-95.
775 Jensen, J. D., Y. Kim, V. B. DuMont, C. F. Aquadro and C. D. Bustamante, 2005
776   Distinguishing between selective sweeps and demography using DNA
777   polymorphism data. Genetics 170**:** 1401-1410.
778 Jensen, J. D., B. A. Payseur, W. Stephan, C. F. Aquadro, M. Lynch *et al.*, 2019 The
779   importance of the Neutral Theory in 1968 and 50 years on: A response to
780   Kern and Hahn 2018. Evolution 73**:** 111-114.
781 Kern, A. D., and M. W. Hahn, 2018 The Neutral Theory in Light of Natural
782   Selection. Molecular Biology and Evolution 35**:** 1366-1371.

783 Kimura, M., 1979 Model of Effectively Neutral Mutations in Which Selective
784   Constraint Is Incorporated. Proceedings of the National Academy of
785   Sciences of the United States of America 76: 3440-3444.
786 Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cabridge, UK:
787   Cambridge Univ. Press.
788 Kimura, M., and T. Ohta, 1971 Protein Polymorphism as a Phase of Molecular
789   Evolution. Nature 229: 467-&.
790 Koenig, D., J. Hagmann, R. Li, F. Bemm, T. Slotte *et al.*, 2018 Long-term balancing
791   selection drives evolution of immunity genes in Capsella. bioRxiv.
792 Kreitman, M., 1996 The neutral theory is dead. Long live the neutral theory.
793   Bioessays 18: 678-683.
794 Li, H., and R. Durbin, 2010 Fast and accurate long-read alignment with Burrows-
795   Wheeler transform. Bioinformatics 26: 589-595.
796 Mace, E. S., S. S. Tai, E. K. Gilding, Y. H. Li, P. J. Prentis *et al.*, 2013 Whole-genome
797   sequencing reveals untapped genetic potential in Africa's indigenous
798   cereal crop sorghum. Nature Communications 4.
799 Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters *et al.*, 2013
800   Genome-wide evidence for speciation with gene flow in Heliconius
801   butterflies. Genome Research 23: 1817-1828.
802 McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The
803   Genome Analysis Toolkit: A MapReduce framework for analyzing next-
804   generation DNA sequencing data. Genome Research 20: 1297-1303.
805 Murray, G.G.R., A. E. R. Soares, B. J. Novak, N. K. Schaefer, J. A. Cahill, A. J. Baker, J.
806   R. Demboski, A. Doll, R. R. Da Fonseca, T. L. Fulton, M. T. P. Gilbert, P. D.
807   Heintzman, B. Letts, G. McIntosh, B.L. O'Connell, M. Peck, M.-L. Pipes, E. S.
808   Rice, K. M. Santos, A. G. Sohrweide, S. H. Vohr, R. B. Corbett-Detig, R. E.
809   Green and B. Shapiro 2017 Natural selection shaped the rise and fall of
810   passenger pigeon genomic diversity. Science *358*(6365), 951–954.
811 Nei, M., Y. Suzuki and M. Nozawa, 2010 The Neutral Theory of Molecular
812   Evolution in the Genomic Era. Annual Review of Genomics and Human
813   Genetics, Vol 11 11: 265-289.
814 Novikova, P. Y., N. Hohmann, V. Nizhynska, T. Tsuchimatsu, J. Ali *et al.*, 2016
815   Sequencing of the genus Arabidopsis identifies a complex history of
816   nonbifurcating speciation and abundant trans-specific polymorphism.
817   Nature Genetics 48: 1077-+.
818 Ohta, T., 1972 Population Size and Rate of Evolution. Journal of Molecular
819   Evolution 1: 305-314.
820 Ohta, T., 1973 Slightly Deleterious Mutant Substitutions in Evolution. Nature
821   246: 96-98.
822 Ohta,T., 1977 Extension to the neutral mutation random drift hypothesis, pp.
823   148-167 in *Molecular Evolution and Polymorphism,* edited by M. Kimura
824   National Institute of Genetics, Mishima, Japan.
825 Ohta, T., 1992 The Nearly Neutral Theory of Molecular Evolution. Annual Review
826   of Ecology and Systematics 23: 263-286.
827 Ohta, T., and J. H. Gillespie, 1996 Development of neutral and nearly neutral
828   theories. Theoretical Population Biology 49: 128-142.
829 Pavlidis, P., and N. Alachiotis, 2017 A survey of methods and tools to detect
830   recent and strong positive selection. Journal of Biological Research-
831   Thessaloniki 24.

832   Posada, D., and T. R. Buckley, 2004 Model selection and model averaging in
833   phylogenetics: Advantages of akaike information criterion and Bayesian
834   approaches over likelihood ratio tests. Systematic Biology 53**:** 793-808.
835   R Core Team, 2018 R: A language and environment for statistical computing. R
836   Foundation for Statistical Computing, pp. R Foundation for Statistical
837   Computing, Vienna, Austria.
838   Rousselle, M., M. Mollion, B. Nabholz, T. Bataillon, and N. Galtier, 2018
839   Overestimation of the adaptive substitution rate in fluctuating
840   populations. *Biology Letters*, *14*(5).
841   Sella, G., D.A. Petrov, M. Przeworski and P. Andolfatto 2009 Pervasive natural
842   selection in the Drosophila genome? PLoS genetics 5: e1000495
843   Sjödin, P., I. Kaj, S. Krone, M. Lascoux, and M. Nordborg. 2005. On the meaning and
844   existence of an effective population size. Genetics 169:1061-1070.
845   Stanley, C. E., and R. J. Kulathinal, 2016 Genomic signatures of domestication on
846   neurogenetic genes in Drosophila melanogaster. Bmc Evolutionary
847   Biology 16.
848   Tajima, F., 1989 Statistical-Method for Testing the Neutral Mutation Hypothesis
849   by DNA Polymorphism. Genetics 123**:** 585-595.
850   Tataru, P., M. Mollion, S. Glemin and T. Bataillon, 2017 Inference of Distribution
851   of Fitness Effects and Proportion of Adaptive Substitutions from
852   Polymorphism Data. Genetics 207**:** 1103-1119.
853   Torres, R., M.G. Stetter, R.D. Hernandez and J. Ross-Ibarra, 2019 The temporal
854   dynamics of background selection in non-equilibrium populations.
855   BioRxiv **doi:** https://doi.org/10.1101/618389
856   Vigué L, Eyre-Walker A. 2019 The comparative population genetics of
857   *Neisseria meningitidis* and *Neisseria gonorrhoeae.* PeerJ 7:e7216
858   Walsh, B., and M. Lynch, 2018 *Evolution and Selection of Quantitative Traits*.
859   Oxford University Press   .
860   Weissman, D. B., and N. H. Barton, 2012 Limits to the Rate of Adaptive
861   Substitution in Sexual Populations. Plos Genetics 8.
862   Welch, J. J., A. Eyre-Walker and D. Waxman, 2008 Divergence and Polymorphism
863   Under the Nearly Neutral Theory of Molecular Evolution. Journal of
864   Molecular Evolution 67**:** 418-426.
865   Zeng, K. 2013. A coalescent model of background selection with recombination,
866   demography and variation in selection coefficients. Heredity 110: 363-
867   371
868
869

870

871

APPENDIX

873

874 In a constant population with population size N_e, $\pi_S = 4N_e\mu$ and π_N is given by

875 (Sawyer and Hartl 1992):

876
$$\pi_N = 2N_e\mu \int_0^1 2x(1-x)H(S,x)dx \quad (A1)$$

877 where

878
$$H(S,x) = \frac{1-e^{-S(1-x)}}{x(1-x)(1-e^{-S})} \qquad (A2)$$

879 is the mean time a new semidominant mutation of scaled selection coefficient $S =$

880 $4N_e s$ spends between x and $x + dx$ (Wright 1938). For constant selection S, by

881 integrating (A1) and dividing by $4N_e\mu$, we have:

882
$$\frac{\pi_N}{\pi_S} = f(S) = \frac{2}{1-e^{-S}} - \frac{2}{S} \qquad (A3)$$

883 (A3) is valid for both positive and negative fitness effect. If we consider only

884 beneficial mutations with a gamma distribution of effects, with mean S_b and

885 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

886 as Welch et al. (2008) to show that:

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

887
$$=\frac{1}{\beta_b-1}\left(\frac{\beta_b}{S_b}\right)^{\beta_b}\left(\xi\left(\beta_b-1,\frac{\beta_b}{S_b}+1\right) + (\beta_b-1)\xi\left(\beta_b,\frac{\beta_b}{S_b}\right) - \xi\left(\beta_b-1,\frac{\beta_b}{S_b}\right)\right) \ (A4)$$

888 where $\xi(x,y)$ is the Hurwith Zeta function. (A4) can be approximated under the

889 realistic assumption that $\frac{\beta_b}{S_b} \ll 1$ and taking Taylor expansion of (A4) in $\frac{\beta_b}{S_b}$

890 around 0. We thus obtain:

891 $\frac{\pi_N}{\pi_S} \approx (2\pi)^{\beta_b}\left(\frac{S_b}{\beta_b}\right)^{\beta_b}$ (A5)

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

893

894

895

896 **Table 1** Species and datasets used in the present study

897

Species	Ref.	Outgroup	Ref.	Mating type	AIC	b	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

898 Note: AIC values were estimated by *polyDFE* for models with and without the effects of beneficial mutations, respectively (bold numbers showed significance <

899 0.05). The same applies to β_{full} and β_{gamma} as well. β_{full} and β_{gamma} were the shape parameter for full DFE and gamma DFE model respectively. β_{max} corresponds to the

900 maximum value of those estimated by *polyDFE* for each ranked gene bin.

901

902 **Table 2** Summary table of multiple regression analyses of the effects of π_4, S_b,

903 Tajima's D, and ρ_D on Δ, the difference between b and β.

904

$\Delta \sim \pi_4 + \log_{10}(S_b)$	*Coef.*	*SE*	*t value*	*p-value*
Intercept	0.14	0.031	4.69	0.0016**
π_4	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			
$\Delta \sim \pi_4 + D + \rho_D$				
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_4	15.80	3.39	4.65	0.0040**
p-value: 0.002978	Adjusted R^2: 0.708			

905
906 ***: p<0.001, **: 0.001<p<0.01, *: 0.01<p<0.05, ˙: 0.05<p<0.1

907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932

933

934 **Table 3** Changes of summary statistics and DFE parameters across 20 rank gene

935 groups.

936

| | Tajima's D | | $\rho_\beta{}^a$ | $\rho p_b{}^a$ |
	median	$\rho_D{}^a$		
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•
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

937

938 a: ρ is the slope of the regression of D (β, and p_b, respectively) over genetic

939 diversity across ranked groups of genes.

940 ***: $p<0.001$, **: $0.001<p<0.01$, *: $0.01<p<0.05$, •: $0.05<p<0.1$

941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963

964 **Table 4** Results of forward simulations showing the effect of linked positive selection on b, Δ and summary statistics of the site frequency
965 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
966 π_S and Tajima's D.

|	N	S_b	S_d	β	b	Δ	π_4	π_0/π_4	ρ_D	D
Nr=0	100	20	1000	0.4	0.49	0.09	1.39	0.091	874.6	-0.46
	100	50	1000	0.4	0.61	0.21	1.18	0.094	909.9	-0.70
	100	100	1000	0.4	0.72	0.32	1.06	0.111	994.2	-0.77
	100	10	1000	0.4	0.46	0.06	1.52	0.082	739.9	-0.36
	500	10	1000	0.4	0.65	0.25	5.72	0.09	228.6	-0.77
	1000	10	1000	0.4	0.81	0.41	10.35	0.094	132.4	-0.92
Nr=1e-3	100	20	1000	0.4	0.06	-0.34	1.64	0.076	662.5	-0.18
	100	50	1000	0.4	0.63	0.23	1.48	0.087	738.1	-0.28
	100	100	1000	0.4	0.72	0.32	1.17	0.097	966.8	-0.58
	100	10	1000	0.4	0.09	0.031	1.70	0.075	1011.1	-0.12
	500	10	1000	0.4	0.61	0.21	7.54	0.084	163.9	-0.26
	1000	10	1000	0.4	0.68	0.28	13.67	0.083	99.7	-0.37
Nr=1e-2	100	20	1000	0.4	0.43	0.03	1.74	0.077	739.3	-0.048
	100	50	1000	0.4	0.63	0.23	1.67	0.081	917.6	-0.12
	100	100	1000	0.4	0.78	0.38	1.61	0.084	898.4	-0.15
	100	10	1000	0.4	0.33	-0.07	1.76	0.080	325.7	-0.011
	500	10	1000	0.4	0.69	0.29	8.55	0.073	165.4	-0.06
	1000	10	1000	0.4	0.99	0.59	16.7	0.072	86.3	-0.12

Figures
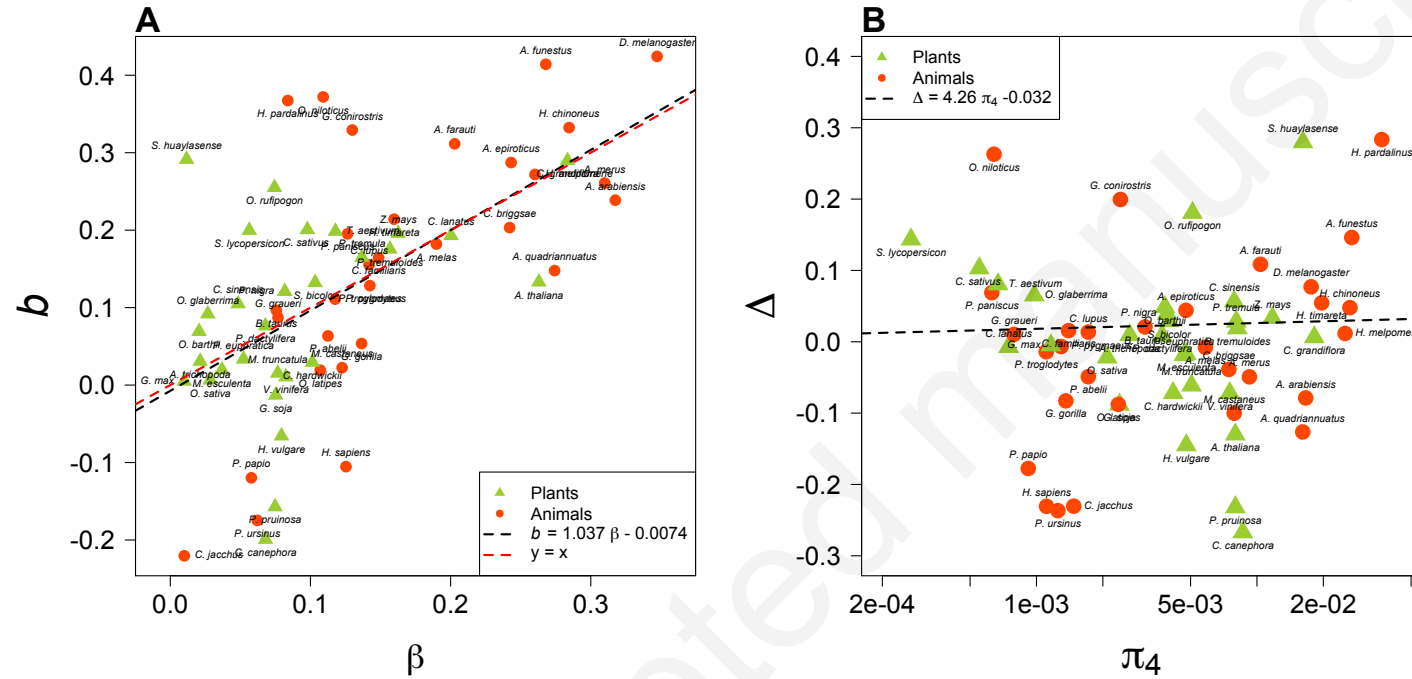
**Fig. 1** (A) The correlation between *b* and the shape parameter of the DFE, *β,* from the 59 species in Chen *et al.* (2017). The observed slope of the regression of $\log(\pi_0/\pi_4)$ over $\log(\pi_4)$, *l=-b*. (B) The distribution of Δ (=*b-β*) against genetic diversity at synonymous sites. *β* values were estimated from DFE models with only deleterious mutations considered (the gamma distribution).

975
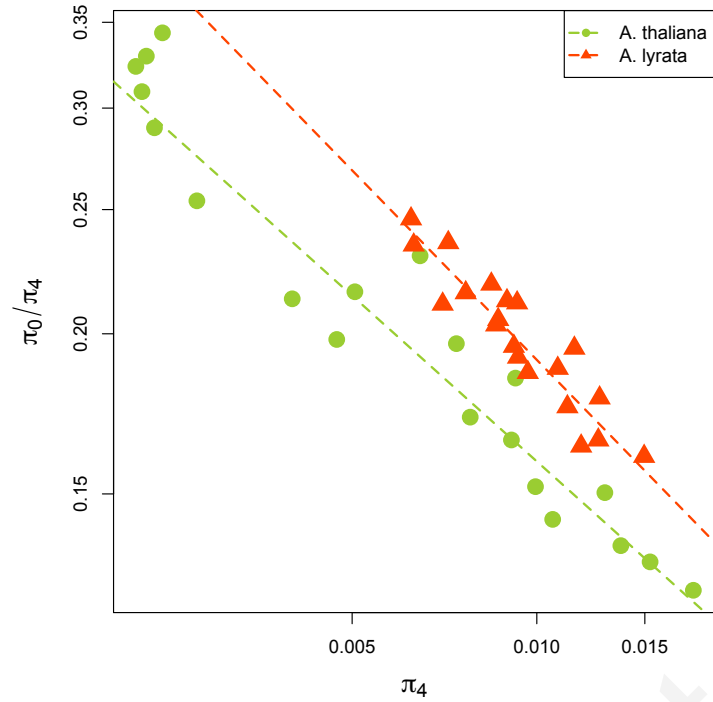
Fig. 2 The regression of $\log(\pi_0/\pi_4)$ over $\log(\pi_4)$ for self-fertilizing *Arabidopsis thaliana* (dots) and its outcrossing relative *A. lyrata* (triangles).

978
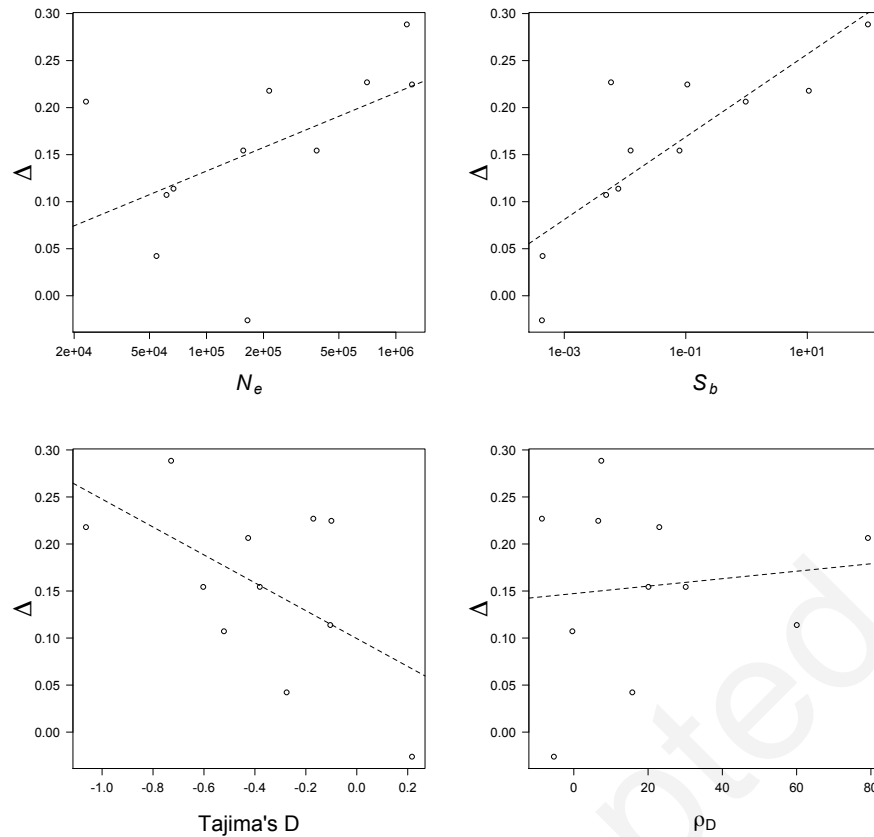
Fig. 3 The relationship between Δ (=*b*-*β*) and effective population size, N_e, selective strength, S_b, Tajima's D and the trend of D across bins ρ_D for 11 selected species. Dotted lines showed the linear regression line. *β* and S_b values were estimated from full DFE models with both deleterious and beneficial mutations considered (full DFE model with both gamma and exponential distributions).

982

983
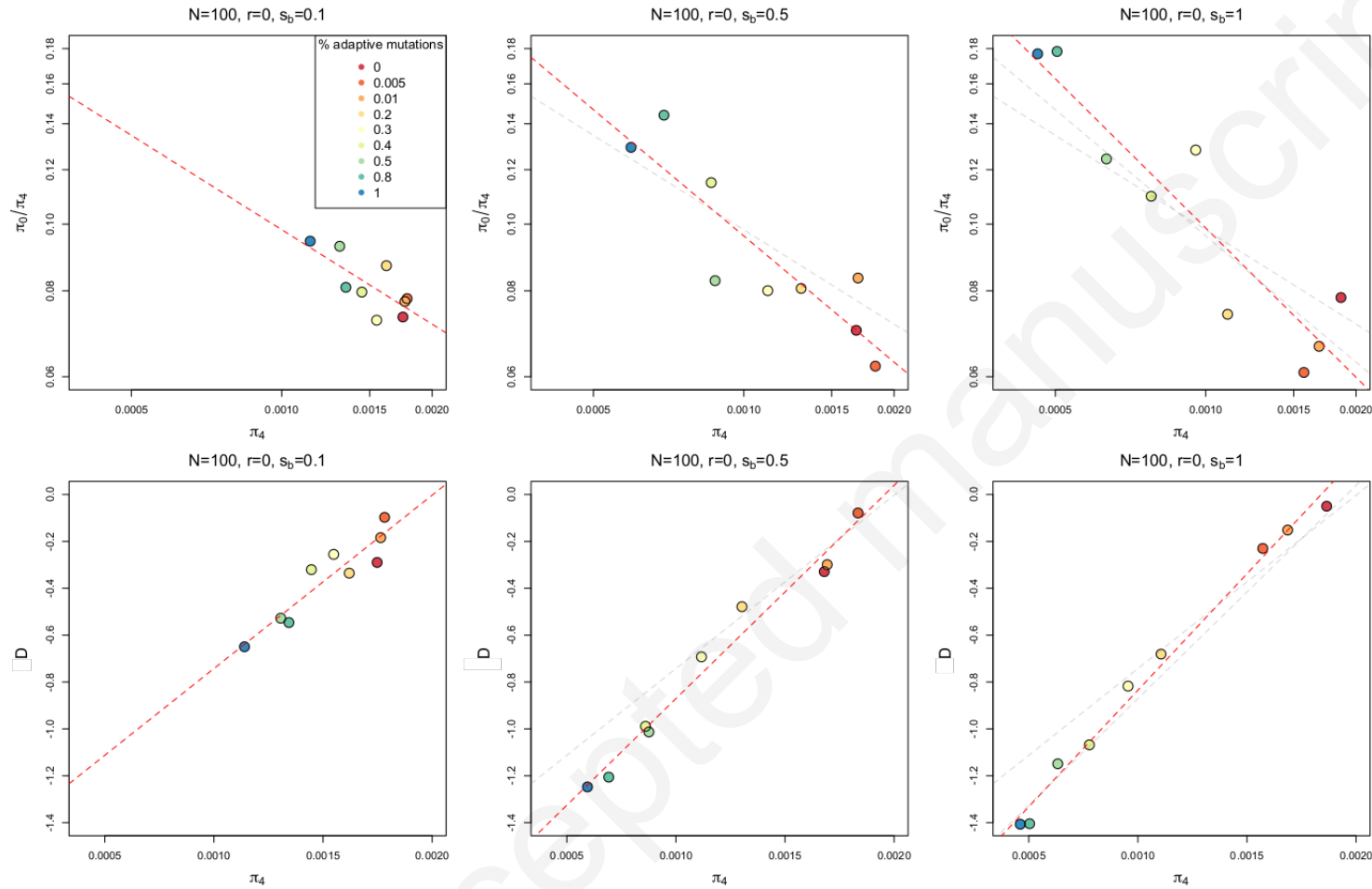
Fig. 4 Effect of linked positive selection on the relationship between $\log(\pi_0/\pi_4)$ and $\log(N_e)$ and Tajima's D. Upper row: The linear

regression coefficient (*b*) between $\log(\pi_0/\pi_4)$ and $\log(N_e)$ increases with increasing positive selective strength (from left to right). The

986   red lines are the regression lines for each case. To facilitate comparisons among figures, and illustrate how the slope gets steeper as s_b

987   increases the regression lines corresponding to s_b=0.1 and/or s_b=0.5 values are reported with gray lines. Lower row: The red lines for

988   Tajima's D panels indicate the mean values.

989
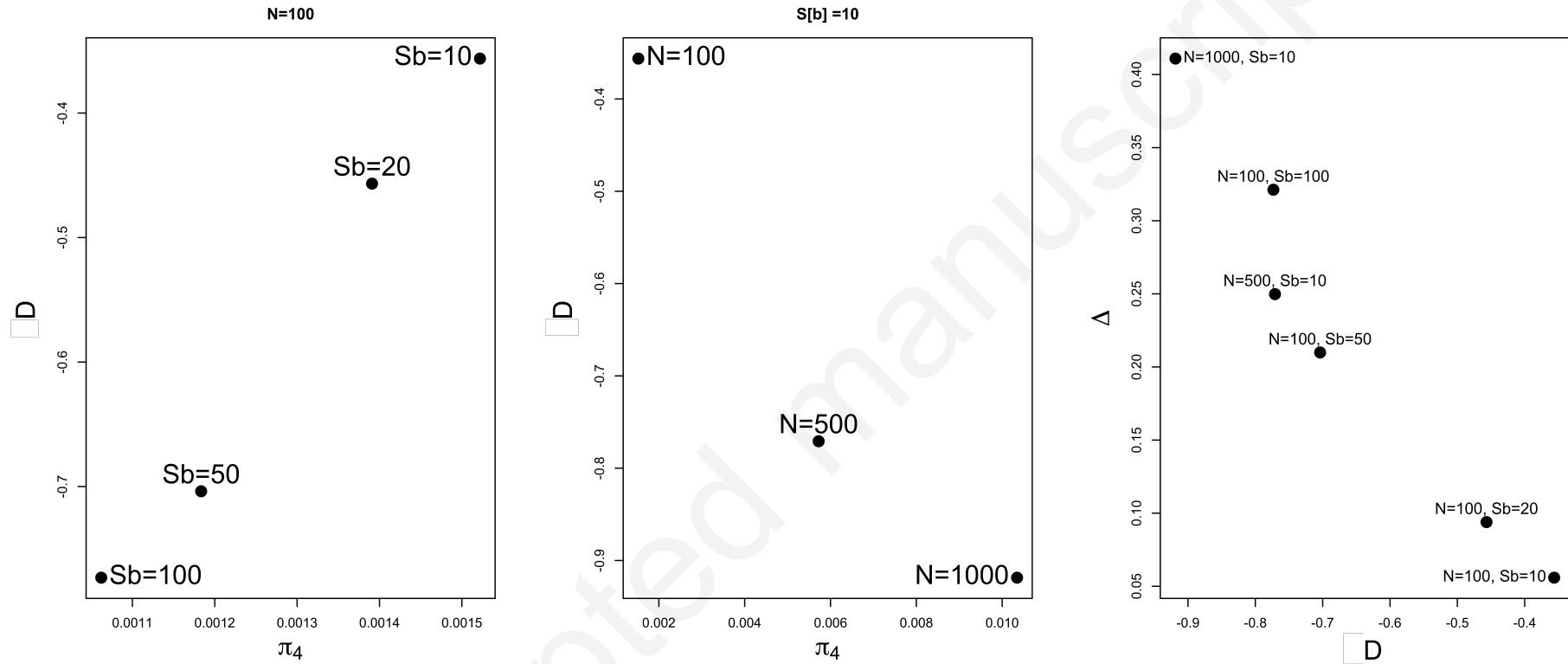
Fig. 5 The correlation between Tajima's D and π_4 depending on S_b (left panel) and N (middle panel); Correlation between Δ and Tajima's D (right panel). In all three cases the results were obtained with forward simulations in Slim assuming no recombination.

Supplementary Information

Supplementary table

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. Mutation rates used for 11 species used in the current study for estimation of N_e.

Table S4. 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	#chromosom	#genes	slope (I)	R2	p.value	S_d	beta	slope(-5)	R2(-5)	p-value(-5)
A. trichopoda	8	9002	-0.03079	0.0122	0.4706823	1.85e+22	0.021322045	-0.14312	0.1152	0.1168
A.thaliana	20	14308	-0.13308	0.4698	0.00382	208.0715692	0.262947282	-0.20209	0.7542	1.637e-05
S.bicolor	7	12382	-0.13228	0.5473	0.00116	273522.5492	0.10331407	-0.15113	0.6233	0.0002819
M.truncatula	20	7822	-0.015393	0.0081939	0.51873	3520000	0.076671273	-0.01805	-0.06365	0.6937
P.nigra	18	8009	-0.12091	0.19883	0.102593	7510000	0.081667091	-0.11810	0.1453	0.0889
P.tremula	20	17530	-0.165	0.6679	2.92e-05	27081.48075	0.13677788	-0.14738	0.4328	0.00458
P.tremuloides	20	16777	-0.1756	0.6351	4.4e-05	9513.471827	0.156690454	-0.14800	0.6423	0.000199
P.euphratica	40	12739	-0.033542	0.03494	0.3758	3,00E+08	0.052221393	-0.09721	0.2191	0.04483
P.pruinosa	40	15872	0.15714	0.5765	0.000812	1740000	0.074766962	0.005517	0.9326	-0.07631
V.vinifera	20	10029	-0.010585	-0.002714	0.53938	2920000	0.082468092	-0.01496	-0.07498	0.8805
T. aestivum	5	13135	-0.1985	0.6614	4,00E-04	13631.07439	0.117827362	-0.11301	0.2179	0.04535
C.sativus	19	8107	-0.2008	0.37419	0.0346859	3760000	0.097784926	-0.15807	0.3058	0.019
C.hardwickii	10	8075	-0.02948	0.005111	0.52121	4390000	0.101323245	-0.07783	-0.02329	0.424
Z. mays	10	1676	-0.1959	0.292498	0.0379831	18769.08991	0.162600552	-0.16579	0.2745	0.02612
G.soja	20	23902	0.01296	0.00965	0.4587	5070000	0.075207188	-0.1847	0.08996	0.1466
G.max	20	23721	-0.005101	0.004733	0.43566	5.35e+48	0.01	-0.05874	-0.04836	0.562
C.sinensis	4	10983	-0.10496	0.38487	0.0164	1.93e+09	0.048424245	-0.09091	0.2632	0.0292
O.sativa	20	12416	-0.00658	-0.02686	0.6038	9.12e+13	0.028516702	-0.01644	-0.0603	0.6591
O.rufipogon	11	6305	-0.2551	0.6121	0.001	3190000	0.074351324	-0.31151	0.4939	0.00209
O.glab	13	8849	-0.09186	0.37368	0.0234	8.11e+16	0.026769072	-0.06629	0.3779	0.00873
O.barthii	9	6133	-0.06925	0.16368	0.1226006	8.94e+23	0.02054373	-0.11572	0.05067	0.20901
C.canephora	7	11528	0.1991	0.5222	0.00106	5640000	0.067941908	0.08392	0.1381	0.09499
C. lanatus	10	6038	-0.19304	0.18761	0.1244345	357.9045775	0.200255858	-0.25481	0.3883	0.007759
M.esculenta	14	12536	-0.01945	0.025044	0.386128	1.69e+15	0.036875397	-0.03299	-0.01368	0.384
H.vulgare	4	6232	0.06568	0.12179	0.1669793	4540000	0.079329761	-0.01498	-0.06763	0.7419
C. grandiflora	20	12667	-0.2898	0.8196	3.39e-07	954.9860899	0.283390544	-0.27222	0.7204	3.85e-05
P.dactylifera	20	14166	-0.07643	0.240101	0.0538646	20400000	0.067902271	-0.03901	-0.01052	0.3722
S.lycopersicon	5	14665	-0.1998	0.000184	0.6199	43400000	0.056347066	-0.14925	0.4151	0.00567
S.huaylasense	6	14684	-0.2914	0.000216	0.6211	5.97e+53	0.011645271	-0.4157	0.4673	0.00297
H.sapiens	20	18191	0.105273	0.046502	0.38661	10002.76568	0.12535229	-0.2185	0.09951	0.1345
P.troglodytes	20	16333	-0.12841	0.08777	0.28757	9517.764368	0.142391471	-0.3956	0.3142	0.01741
P.paniscus	20	15233	-0.195248	0.12181	0.2393388	9478.600653	0.126442144	-0.2271	0.1624	0.0761
G. gorilla	20	12348	-0.053642	0.028298	0.412998	6536.50279	0.136471801	-0.15955	0.2861	0.02325
G. graueri	6	13334	-0.087203	0.09082	0.326128	3920000	0.076851226	-0.20960	0.2629	0.0293
P. abelii	10	15925	-0.06348	0.001645	0.49322	176632.7316	0.112477663	-0.31915	0.7538	1.655e-05
P. pygmaeus	10	15570	-0.11097	0.08732	0.263535	49157.40682	0.117581567	-0.27782	0.5856	0.0005365
P. papio	4	13335	0.11973	0.3219	0.0179	66300000	0.057833203	-0.08248	-0.02491	0.4313
P. ursinus	4	13283	0.1747	0.4082	0.00409	69900000	0.062022659	0.04078	-0.06097	0.6656
C. jacchus	10	12859	0.2204	0.5907	0.000123	1.09e+44	0.00999999	0.1148	0.0009799	0.3324
C. familiaris	20	12670	-0.16431	0.125731	0.2395	58538.91338	0.148629986	-0.2936	0.2217	0.04371
C. lupus	8	12665	-0.1555	0.11165	0.199697	123933.2123	0.141824095	-0.35445	0.4613	0.00321
B. taurus	18	13714	-0.09646	0.02092	0.37583	1.94e+09	0.075685116	-0.2152	0.1356	0.097166
O. latipes	20	5478	-0.01919	0.02083	0.4351	1290000	0.106989872	-0.06800	-0.03471	0.4794
O. niloticus	6	4939	-0.372	0.08717	0.1503	389319.476	0.109072768	-0.2716	-0.008202	0.3637
G. conirostris	20	2142	-0.3293	0.136386	0.233182	2990000	0.129853019	-0.3212	0.1295	0.1026
D.melanogaster	20	3686	-0.4243	0.7958	3.94e-06	3974.566141	0.347297266	-0.40317	0.7988	4.37e-06
C. briggsae	10	2497	-0.20332	0.18928	0.1225722	435.2635717	0.242110286	-0.26570	0.382	0.00833
M. castaneus	20	19126	-0.022442	0.006106	0.48217	1890000	0.122589716	-0.15436	0.2057	0.05091
A. arabiensis	20	6763	-0.2387	0.6917	7.67e-05	3966.393817	0.317476941	-0.26315	0.5852	0.00054
A. epiroticus	20	6558	-0.2873	0.5709	0.00118	3844.612974	0.243166406	-0.34489	0.5538	0.0008843
A. farauti	20	6264	-0.3115	0.4937	0.00448	27491.72162	0.202916231	-0.37540	0.6613	0.000138
A. funestus	12	6867	-0.4141	0.7947	5.43e-07	35258.79598	0.267911067	-0.38183	0.74	2.38e-05
A. melas	12	7148	-0.1821	0.30286	0.0201097	29934.62801	0.189863622	-0.02452	-0.07275	0.8256
A. merus	20	6665	-0.2608	0.6139	0.000194	2218.056088	0.309945788	-0.2981	0.567	0.0007221
A. quadriannuatus	20	6620	-0.1478	0.4223	0.0117	10813.37126	0.274193762	-0.19368	0.5152	0.001555
H. melpomene	8	6567	-0.2719	0.7328	4.29e-06	15840.97789	0.26019666	-0.33241	0.836	1.133e-06
H. chinoneus	8	6437	-0.3324	0.7134	1.83e-05	9256.968993	0.284517894	-0.44241	0.8041	3.66e-06
H. timareta	8	6434	-0.2142	0.6244	0.000243	1790000	0.159669111	-0.21311	0.5176	0.0015
H. pardalinus	4	6459	-0.3673	0.6618	7.2e-05	1.49e+12	0.083838116	-0.40164	0.6631	0.000133

note: slope (-5), R2(-5), p-value(-5) are correlation results of log(piN/piS) and log(piS) with five highest piS bins removed

Species	filtering*	#chromosomes	SNPs_used (failed in polarization)	b	CI_boots(95%)	p-value	R2	r2_boots(95%)	beta	S_d	p_b	S_b	pi0/pi4	Pi0	Pi4	mutation	TD	rhoD
A. thaliana	10, 0, 0, 1, 100, 0.1	10	209836 (5.40%)	0.477	(0.435, 0.521)	3.60E-07	0.975	(0.938, 0.995)	0.3225751	-222.003	7.98E-06	0.0123487	0.233257195	0.0010212	0.004378	7	-0.38078969	20.09847
A. lyrata	10, 0, 0, 1, 100, 0.1	10	290964 (12.53%)	0.499	(0.585, 0.41)	4.02E-08	0.88	(0.765, 0.947)	0.3446372	-352.4657	1.95E-05	0.0787558	0.183629727	0.0019619	0.010684	7	-0.60281367	30.13018
C. rubella	10, 0, 0,30,100,0.5	10	101053 (0.87%)	0.43	(0.387, 0.466)	4.78E-11	0.953	(0.899,0.98)	0.3877494	-280.1572	0.05308512	0.000439121	0.244241748	0.0003722	0.0015239	7	-0.27545110	15.75497946
C. grandiflora	10, 0, 0, 1, 100, 0.5	16	185372 (0.33%)	0.521	(0.359, 0.695)	1.20E-04	0.677	(0.457, 0.839)	0.3030656	-645.0785	0.0117206	10.58446	0.2	0.0012	0.006	7	-1.06321752	23.02078179
S. habrochaites	1e-20, 200, 0, 1, 100, 0.1	7	329695 (0.0063%)	0.205	(0.08, 0.319)	0.0207338	0.381	(0.0495, 0.69)	0.2311101	-98565.41	0.1430495	0.000428568	0.203418054	0.0006963	0.003423	5.2	0.216722329	-5.36317782
S. huaylasense	1e-20, 200, 0, 1, 100, 0.1	4	104101 (0.10%)	0.536	(0.423, 0.656)	3.76E-06	0.79	(0.612, 0.9)	0.309072	-78719.78	0.1444625	0.005878293	0.175546448	0.00257	0.01464	5.2	-0.17066493	-8.58607658
S. propinquum	10, 0, 0, 1, 10, 0.1	7	129469 (0.52%)	0.374	(0.32, 0.426)	1.00E-08	0.904	(0.816, 0.959)	0.260165	-284.4627	2.93E-06	0.007746282	0.252857677	0.0006769	0.002677	10	-0.10394778	60.03524
Z. mays	10, 0, 0, 1, 100, 1	10	179092 (29.3%)	0.292	(0.262, 0.319)	1.30E-10	0.942	(0.893, 0.972)	0.1848003	-2525.62	1.86E-05	0.004883847	0.331255083	0.002444	0.007378	30	-0.52155749	-0.38718011
P. trichocarpa	10, 0, 0, 1, 10, 0.1	16	575576 (0)	0.421	(0.277, 0.598)	1.56E-04	0.679	(0.43, 0.853)	0.2146055	-5353.714	0.03788233	0.9679363	0.220240157	0.000763	0.0034644	37.5	-0.42651297	79.19642
D. melanogaster	10, 0, 0, 1, 10, 0.1	20	497276 (1.45%)	0.7	(0.62, 0.768)	1.26E-11	0.95	(0.909, 0.978)	0.41152	-2175.355	0.00831	99.45153	0.090948175	0.0011635	0.012793	2.8	-0.72904240	7.406745777
H. timareta	10, 0, 0, 1, 100, 0.1	8	190601 (0.1%)	0.435	(0.39, 0.476)	1.60E-09	0.914	(0.845, 0.961)	0.2103758	-94578.76	0.001163899	0.1059159	0.109142452	0.00154	0.01411	2.9	-0.09977894	6.583048719

*: for filtering we performed following criteria in order: e-value, bit-score, query coverage, query length, num of low quality sites, and percentage of low quality sites

different filtering criteria were chose for each species in order to maximize the linearity (R2 column) for slope calculation