

Quantifying the variation in the effective population size within a genome.

Research article

Toni I. Gossmann¹, Megan Woolfit², Adam Eyre-Walker^{1,*}

¹ School of Life Sciences, University of Sussex, Brighton, United Kingdom

² School of Biological Sciences, University of Queensland, Brisbane, Australia

*Correspondence: a.c.eyre-walker@sussex.ac.uk

September 24, 2011

Abstract

The effective population size (N_e) is one of the most fundamental parameters in population genetics. It is thought to vary across the genome as a consequence of differences in the rate of recombination and the density of selected sites due to the processes of genetic hitch-hiking and background selection. Although it is known that there is intragenomic variation in the effective population size in some species, it is not known whether this is widespread, nor how much variation in the effective population size there is. Here, we test whether the effective population size varies across the genome, between protein coding genes, in 10 eukaryotic species by considering whether there is significant variation in neutral diversity, taking into account differences in the mutation rate between loci by using the divergence between species. In most species we find significant evidence of variation. We investigate whether the variation in N_e is correlated to recombination rate and the

density of selected sites in four species, for which this data is available. We find that N_e is positively correlated to recombination rate in one species, *Drosophila melanogaster* and negatively correlated to a measure of the density of selected sites in two others, humans and *Arabidopsis thaliana*. However, much of the variation remains unexplained. We then use a hierarchical Bayesian analysis to quantify the amount of variation in the effective population size and show that it is quite modest in all species - most genes have an N_e which is within a few fold of all other genes. Nonetheless we show that this modest variation in N_e is sufficient to cause significant differences in the efficiency of natural selection across the genome, by demonstrating that the ratio of the number of non-synonymous to synonymous polymorphisms is significantly correlated to synonymous diversity and estimates of N_e , even taking into account the obvious non-independence between these measures.

Introduction

The effective population size (N_e) is one of the most fundamental quantities in population genetics, evolutionary biology and molecular ecology, since it determines the effectiveness of natural selection and the level of neutral genetic diversity that a population contains (CHARLESWORTH, 2009). Populations and regions of the genome with small N_e tend to have low levels of genetic diversity, to be susceptible to the accumulation of deleterious mutations through genetic drift and to have potentially low rates of adaptive evolution (CHARLESWORTH, 2009).

The effective population size is expected to vary across the genome as a consequence genetic hitch-hiking (SMITH and HAIGH, 1974) and background selection (CHARLESWORTH *et al.*, 1993). The action of both positive and negative natural selection, particularly in regions of the genome with low rates of recombination, is expected to reduce the effective population size leading to lower levels of genetic diversity and reduced effectiveness of selection. Hence variation in the rate of recombination and the density of selected sites is expected to generate variation in N_e .

The evidence that there is variation in N_e within a genome comes from three sources. First, it has been shown that levels of neutral genetic diversity are correlated to rates of recombination in *Drosophila* (BEGUN and AQUADRO, 1992), humans (HELLMANN *et al.*, 2003) and some plant species (TENAILLON *et al.*, 2004; ROSELIUS *et al.*, 2005). This could be due to variation in the mutation rate since neutral genetic diversity is proportional to the effective population size multiplied by the mutation rate. However, the level of neutral sequence divergence between species, which should be proportional to the mutation rate, is not correlated to the rate of recombination in *Drosophila* (BEGUN and AQUADRO, 1992) and the plant species (ROSELIUS *et al.*, 2005) that have been investigated. Furthermore, although there is a correlation between neutral sequence divergence and recombination rate in humans, this correlation is not sufficient to explain the correlation between diversity and the recombination rate (HELLMANN *et al.*, 2005). It is has also been shown that the Y and W chromosomes, which have no recombination over most of their length, have substantially lower diversity than other chromosomes, and that this cannot be attributed to differences in the mutation rate or the fact there are fewer Y and W chromosomes than autosomes (MONTELL *et al.*, 2001; FILATOV *et al.*, 2001; BACHTROG and CHARLESWORTH, 2002; HELLBORG and ELLEGREN, 2004). It thus seems that the effective population size varies across genomes and is positively correlated to the rate of recombination.

Second, under the neutral theory of molecular evolution it is expected that levels of diversity and divergence should be proportional to each other, since both depend on the neutral mutation rate. Deviations from this hypothesis, caused for by variation in N_e , can be tested using the HKA test and derivatives of it (HUDSON *et al.*, 1987; WRIGHT and CHARLESWORTH, 2004; INGVARSSON, 2004; INNAN, 2006). Evidence for departures from the neutral hypothesis, based on the HKA test, comes from multiple multilocus surveys in plants (ROSELIUS *et al.*, 2005; SCHMID *et al.*, 2005), the chicken Z chromosome (SUNDSTRÖM *et al.*, 2004), humans (ZHANG *et al.*, 2002) and *Drosophila* (MORIYAMA and POWELL, 1996; MACHADO *et al.*, 2002).

Third, variation in the effective population size should manifest itself as variation in the

effectiveness of selection and this has also been observed. In *Drosophila* it has been shown that codon usage bias is lower in the regions of the genome with very low rates of recombination (HEY and KLIMAN, 2002; MARAIS *et al.*, 2003; KLIMAN and HEY, 2003). It has also been shown that the number of non-synonymous polymorphisms (P_n) relative to the number of synonymous polymorphisms (P_s) is higher in the low recombining parts of the *D. melanogaster* genome (PRESGRAVES, 2005), that the rate of non-synonymous (d_N) relative to the rate of synonymous (d_S) substitution is positively correlated to the frequency of recombination (BETANCOURT and PRESGRAVES, 2002) and that the overall efficiency of selection appears to be lower in the regions of the genome with low rates of recombination (PRESGRAVES, 2005; LARRACUENTE *et al.*, 2008). Likewise it has been shown that d_N/d_S is higher on the Y or W chromosome than on the other chromosomes in humans (WYCKOFF *et al.*, 2002) and birds (BERLIN and ELLEGREN, 2006) and the fourth chromosome of *Drosophila* species (ARGUELLO *et al.*, 2010). In contrast, BULLAUGHEY *et al.* (2008) found no correlation between d_N/d_S and the rate of recombination in primates.

It is thought that the correlation between d_N/d_S or P_n/P_s and the rate of recombination is due to regions of the genome with little or no recombination having low effective population size and hence reduced effectiveness of natural selection (BETANCOURT *et al.*, 2009). P_n/P_s is negatively correlated to the rate of recombination because regions with low effective population size allow more slightly deleterious mutations to segregate for longer. In contrast, d_N/d_S can either be positively or negatively correlated to the rate of recombination depending on the prevalence of advantageous mutations. If advantageous mutations are common then regions of the genome with high rates of recombination are expected to evolve faster because they have a higher effective mutation rate, and because selection is effective on a greater proportion of mutations. In contrast, if advantageous mutations are rare then regions of the genome with high rates of recombination may have low values of d_N/d_S because selection against slightly deleterious mutations is more effective.

Although it is well established that N_e varies across the genome in a few species, it is

unclear whether this is true of all species and, more importantly, how much variation in N_e there is and whether this variation results in differences in the effectiveness of selection. Here we test whether there is variation in the effective population size by considering whether there is significant variation in neutral diversity, taking into account that this might be due to variation in the mutation rate by using the divergence between species to control for differences in the mutation rate. We also quantify the variation in N_e . We estimate N_e from the nucleotide diversity at putatively neutral sites, since this is expected to be equal to $4N_e\mu$ in a diploid organism, where N_e is the effective population size and μ is the mutation rate per generation. We use the divergence between two species at neutral sites as an estimate of the mutation rate per generation. Note that since we are comparing loci within a genome they all share the same generation time (unless they are on the sex chromosomes or in the mitochondrial DNA) and so this does not have to be explicitly taken into account. We can therefore estimate the effective population size for each locus. However, although each individual estimate is unbiased, the distribution of these values has a variance that is greater than the true variance because of sampling error; a locus might have a particularly low diversity just by chance, and not because its effective population size is particularly low. To get round this problem we use a hierarchical Bayesian framework to estimate the distribution of N_e across genes taking into account the sampling error associated with both the polymorphism and divergence data.

We test for and investigate the variation in the effective population size in 10 eukaryotic species including humans, *Drosophila melanogaster*, *Arabidopsis thaliana* and *Saccharomyces paradoxus* (Table 1). We find that there is statistically significant variation in N_e across genes, but that it is rather modest in most of the organisms. We also investigate whether variation in N_e within a genome leads to variation in the proportion of effectively neutral mutations, by testing whether the ratio of the number of non-synonymous to synonymous polymorphisms is correlated to the effective population size, in a way which circumnavigates the obvious non-independence between the two variables. We find overall evidence for a correlation between these two parameters and hence conclude that even modest variation in the effective population size is sufficient to generate variation in the

effectiveness of natural selection.

Materials and Methods

Sequence data We obtained data from different plant species, mouse, fruitfly and yeast using publicly available data from Genbank (<http://www.ncbi.nlm.nih.gov/Genbank>). Polymorphism data for *Homo sapiens* were downloaded from Environmental Genome Project (egp.gs.washington.edu) and Seattle SNPs (pga.gs.washington.edu) web-sites and for *Arabidopsis thaliana* from <http://walnut.usc.edu/2010>. The annotated protein-coding genome of *A. thaliana* was obtained from TAIR 8 (<ftp://ftp.arabidopsis.org>), the annotated *Arabidopsis lyrata* genome was obtained from JGI (<http://genome.jgi-psf.org>). The annotated protein-coding genome of *Pan troglodytes*, *Macaca mulatta* and *Rattus norvegicus* were obtained from Ensembl (<http://www.ensembl.org/info/data/ftp/index.html>). The *S.cerevisiae* genome chromosome III was obtained from <http://www.yeastgenome.org>. We restricted our analysis of *D. melanogaster* to data from the Zimbabwe population, from the *S.paradoxus* dataset to the European population and from the human dataset to African populations, since all of these represent the ancestral populations of the three species (GARRIGAN and HAMMER, 2006; STEPHAN and LI, 2007; LITI *et al.*, 2009). Qualitatively similar results were obtained in the three cases when using global data.

Preparation of the data The analysis was performed using protein coding sequences. Coding regions were assigned using protein coding genomic data or if given, taken from the GenBank input files. Sequences were aligned using Clustalw using default parameter values (THOMPSON *et al.*, 1994). The outgroup ortholog was assigned using the best BLAST (ALTSCHUL *et al.*, 1990) hit or, if given, taken from the polymorphism dataset. We only used polymorphism data for which we could assign an outgroup sequence. For all analyses the number of synonymous substitutions and polymorphisms served as the neutral standard. For computational reasons all sites had to have been sampled in the same number of chromosomes within each species; because some loci had been sampled

in more individuals than others and other loci had missing data, we reduced the dataset to a common number of chromosomes by randomly sampling the polymorphisms at each site without replacement. The numbers of synonymous and nonsynonymous sites and substitutions were estimated by randomly selecting one allele from the polymorphism data and comparing it against the outgroup using the F3x4 model implemented in PAML (YANG, 1997) in which codon frequencies are estimated from the nucleotide frequencies at the three codon positions. The proportion of sites estimated by PAML was also used to compute the numbers of synonymous and non-synonymous sites for the polymorphism data. Although, how we choose to define a site can be important in some circumstances (BIERNE and EYRE-WALKER, 2003) this is not likely to be a problem in the current context because we use the same definition for both the divergence and polymorphism data; as such the number of sites effectively cancels out in most of our analyses (however see discussion of selection on synonymous codon bias below). Statistics concerning numbers of loci, numbers of sites as well as polymorphic sites are shown in Table 1.

Testing for variation in diversity and the effective population size We investigated whether there was significant variation in the level of diversity across the genome using two tests. If we assume there is free recombination within and between loci (or no recombination within and between loci) then variation in diversity can be tested using a simple $(2 \times k)$ χ^2 test of independence across the k loci within each species, where for each locus we have the number of sites with a polymorphism and the number of sites without a polymorphism. Note that this test is only valid when the same number of chromosomes have been sampled across all loci. However, some of the variation in diversity between loci might be due to variation in the genealogy if there is limited or no recombination between loci. We therefore applied a variant of the classic HKA test, but we removed the divergence information from the test. The test statistic X^2 is set up as follows:

$$X^2 = \sum_{i=1}^L (P_i - \hat{E}(P_i))^2 / \hat{V}ar(P_i) \quad (1)$$

where $\hat{E}(P_i)$ and $\hat{Var}(P_i)$ are the expected value and variance of the number of segregating polymorphisms, P , in gene i :

$$\hat{E}(P_i) = M_i \theta \sum_{j=1}^{n-1} 1/j \quad (2)$$

$$\hat{Var}(P_i) = \hat{E}(P_i) + (M_i \theta)^2 \sum_{j=1}^{n-1} 1/j^2 \quad (3)$$

with n being the number of alleles, L the number of loci, $\theta = 4N_e\mu$ and M_i the number of sites in gene i . Estimates of θ were obtained by minimizing the value of X^2 . X^2 is expected to be χ^2 distributed with $(L-1)$ degrees of freedom.

Any variation that we detect in diversity might be due to variation in the mutation rate or variation in the local effective population size. We therefore performed two further analyses to investigate whether there was variation in diversity that could not be explained by variation in the mutation rate, as measured by synonymous divergence between species. The first test was a second approximate $(2 \times k)$ χ^2 test of independence, performed as follows. For each locus we have the number of sites used to estimate the level of silent site divergence (L_d), the estimated number of substitutions (D), the number of sites used to estimate silent site diversity (L_p) and the number of sites with a polymorphism (P). Since L_d and L_p can be different we reduced the divergence or polymorphism dataset, whichever was larger, to the size of the other, resampling without replacement the numbers of substitutions or polymorphisms as appropriate; for example if L_d was half L_p , we sampled L_p sites from the divergence data to generate a sub-sample of the substitutions (D') over $L'_d = L_p$ sites. We can then perform a $(2 \times k)$ χ^2 test where the cells for each gene are the number of sites of sites with a substitution (D') and the number of sites with a polymorphism (P'). Note that the dataset will be reduced using this method resulting in a loss of power. Furthermore this test is only approximate because we assume that the number of substitutions is binomially distributed, whereas in fact it has a more complex distribution because of the correction for multiple hits. Some of the expected values can be very small in both χ^2 tests: we therefore checked the p-values from the χ^2 tests by generating the null distribution for the test. This was performed by randomly assigning

polymorphisms and substitutions across the contingency table preserving the marginal totals. We then recalculated the statistic and performed this 1000 times. The p-value was the proportion of such randomly generated values that exceeded the observed value. Generally we found that the p-value from randomisation, and the p-value assuming our test statistics were χ^2 distributed, were similar (Table S1). We therefore present the results from the standard χ^2 test.

This test assumes free recombination between sites within loci and loci (or no recombination between sites and loci). A more conservative test is the classic HKA test which tests for heterogeneity in the ratio of diversity divided by divergence between loci assuming no recombination within loci, but free recombination between loci. We performed the multiple locus HKA test using software provided by J. Hey (<http://genfaculty.rutgers.edu/hey/software#HKA>). To perform this test we had to exclude loci with zero divergence; for most species this constituted a small fraction of the total number of loci. However we had to exclude *S. bicolor* from the analysis because too many loci showed zero divergence.

Recombination and density of selected sites We obtained estimates of recombination rate variation along chromosomes for *A. thaliana* (SINGER *et al.*, 2006), *D. melanogaster* (HEY and KLIMAN, 2002), *H. sapiens* (KONG *et al.*, 2002) and *M. musculus* (DUMONT *et al.*, 2011). Gene density was estimated as the proportion of coding sites in window sizes of 50KB, 500KB and 5MB. Since results are qualitatively similar, we only discuss results for the window size of 500KB. Conservation scores (SIEPEL *et al.*, 2005) were obtained from the UCSC genome browser (<http://genome.ucsc.edu/>) for *D. melanogaster* across 15 species, *H. sapiens* across 17 species and *M. musculus* across 30 species.

Bayesian analysis To estimate the distribution of N_e we used a hierarchical Bayesian analysis in which we estimate the parameters of the distribution of N_e (Figure S1). If we assume that the population size is stationary the expected number of polymorphisms segregating in a sample of n sequences, P_s , and the number of differences between the outgroup and a single sequence from the ingroup, D_s , are

$$P_s = 4\mu L_p N_e \sum_{i=1}^{n-1} 1/i \quad (4)$$

$$D_s = 2\mu t L_d \quad (5)$$

where L_p and L_d are the number of sites which can have a polymorphism or substitution respectively, μ is the nucleotide mutation rate per generation and t is the time of divergence. We are interested in the distribution of N_e . To estimate this distribution we assume that N_e and μ follow a log-normal or a gamma distribution. Assuming free recombination and using equations (4) and (5) above we can write the likelihood of observing \hat{P}_s polymorphisms and \hat{D}_s substitutions

$$L = \prod X(\hat{D}_s, D_s) X(\hat{P}_s, P_s) M(N_e | \sigma_{N_e}) M(\mu | \sigma_\mu) \quad (6)$$

where $X(S, S(x))$ is the Poisson distribution and $M(N_e | \sigma_{N_e})$ is the probability density of the distribution of N_e , and $M(\mu | \sigma_\mu)$ is the probability density of the distribution of the mutation rate; these distributions are parameterised such that the mean is fixed at unity leaving us to estimate the shape parameter. If there is no recombination within a locus then we can rewrite equation (4) as

$$P_s = 4\tau\mu L_p N_e \sum_{i=1}^{n-1} 1/i \quad (7)$$

where τ is the length of the genealogy scaled such that $E[\tau] = 1$. We can rewrite equation (6), and the likelihood then becomes

$$L = \prod X(\hat{D}_s, D_s) X(\hat{P}_s, P_s) M(N_e | \sigma_{N_e}) M(\mu | \sigma_\mu) M(\tau | n) \quad (8)$$

To calculate the probability density distribution $M(\tau | n)$ of genealogy lengths we randomly simulated 10,000 genealogies, scaling them such that the average total length was unity. In theory it is possible to accommodate ancestral polymorphism into the method, however we found that the method rarely gave stable estimates of σ_{N_e} , particularly in the

no recombination model. We therefore concentrated on datasets in which the influence of ancestral polymorphism was likely to be minimal - i.e. in which the average divergence was $> 5\times$ the average of θ_W (Table 1). If we assume that the ancestral N_e of a locus is correlated to the current N_e , we expect ancestral polymorphism to decrease the apparent variation in N_e .

To estimate the posterior distribution of the parameters σ_{N_e} and σ_μ we used a Monte-Carlo Markov chain running the Metropolis-Hastings algorithm (HASTINGS, 1970). Unfortunately because we have very few synonymous polymorphisms per gene this method tends to underestimate the true value of σ_{N_e} . For most datasets this underestimation is small, but it can be large. We therefore estimated the extent of bias by simulating data under a range of parameter values using the actual numbers of sites from the real data such that the expected numbers of polymorphisms and substitutions were equal to the mean values. For example, if we estimated σ_{N_e} to be 0.5 and σ_μ to be 0.1 we simulated data for σ_μ values of 0.1, 0.2 and 0.3 and for σ_{N_e} values between 0.4 and 1.0 in steps of 0.05. For each simulated dataset we estimated σ_{N_e} and using linear regression we inferred the relationship between $\sigma_{N_e}(\text{estimated})$ and $\sigma_{N_e}(\text{true})$. Using this relationship we inferred the true value of σ_{N_e} from the value estimated from the real data (Figures S2 and S3). To obtain a corrected SE we multiplied the observed standard error by the ratio of the corrected estimate of σ_{N_e} divided by the observed estimate of σ_{N_e} . This slightly underestimates the true SE since we have not taken into account the small amount of error associated with estimating the regression line. To test for heterogeneity in σ_{N_e} between species we assumed that the estimate of σ_{N_e} was normally distributed; under this assumption $(\sigma_{N_e} - \bar{\sigma}_{N_e})^2 / \text{var}(\sigma_{N_e})$ is χ^2 distributed with $k-1$ degrees of freedom for k species. $\bar{\sigma}_{N_e}$ was calculated as a weighted average, where the weights were inversely proportional to the variance of the estimate (EYRE-WALKER, 1996).

Variation of efficiency of selection We tested whether the strength of selection on non-synonymous mutations was correlated to the effective population size, which can be seen as testing if the fraction of deleterious mutations varies with N_e . This can be done

by considering the correlation of P_n/P_s and θ_s or P_n/P_s and N_e ($=\theta_s/(4\mu)$), where P_n and P_s are the numbers of non-synonymous and synonymous mutations respectively and N_e values are point estimates from the genetic diversity and mutation rates taken from the literature. However, P_s and θ_s are not independent. We overcome this problem by splitting P_s into two independent values by generating a random hypergeometric variable as follows (PIGANEAU and EYRE-WALKER, 2009; STOLETZKI and EYRE-WALKER, 2011):

$$P_{s1} = \text{Hypergeometric}(P_s, L_s - P_s, 0.5P_s) \quad (9)$$

$$P_{s2} = P_s - P_{s1} \quad (10)$$

One of the P_s values is used to estimate P_n/P_s (see below) and the other one is used to estimate θ_s . There are two further problems to consider with this method, first, P_n/P_s can be an overestimate or underestimate of the true value of P_n/P_s and second the ratio P_n/P_s is undefined if $P_s = 0$. Both of these problems can be overcome by considering the correlation between ψ and θ_s (PIGANEAU and EYRE-WALKER, 2009):

$$\psi = \frac{P_n}{P_s + 1} \quad (11)$$

Hence using our method to split P_s into independent values we have two independent pairs of θ_s and ψ ; we only present results from one pair. Some of the datasets contain relatively little polymorphism which results in substantial variance of ψ . To overcome this problem we sum data across loci. For this we ranked loci according to their neutral diversity obtained from θ_{s2} and binned them into groups of size n (e.g. 2, 4, 8 and 16). For each group average θ_{s2} and corresponding N_{e2} values were calculated. Furthermore, for each group, the sums of P_n and P_{s1} were calculated in order to calculate ψ_1 . Note, that ψ_2 can be obtained in a similar manner, however results were qualitatively comparable and we therefore only show results for ψ_1 vs θ_2 and N_{e2} . Also we only show results for group size 4 because results for group sizes > 2 were similar. The correlations were performed by calculating Spearman's rank correlation and probabilities were combined using the unweighted Z-method (WHITLOCK, 2005).

Results

To investigate variation in the effective population size within genomes, we assembled protein coding sequences from 10 species. The datasets are from six plant species, three animal species and one fungus. The datasets range in size from 66 to 918 loci per species and from 8 to 40 sequences per gene (Table 1). In all analyses we assume that synonymous mutations are neutral.

Variation of diversity and N_e within a genome using χ^2 and HKA tests The level of genetic diversity appears to vary considerably within each genome (Figure 1); however, the number of polymorphisms per gene is generally quite low and hence this variation might be due to sampling error. To test whether the variation is significant we used two tests, which make different assumptions about the rate of recombination within loci - either free or no recombination. Both tests suggest that there is variation in the level of diversity in most species; all species are significant assuming free recombination and 6 out of 10 are significant assuming no recombination. This variation in diversity between loci could be due to variation in the effective population size or to variation in the mutation rate. To investigate whether variation in the mutation rate might be responsible, we estimated the number of synonymous substitutions for each locus (D_S), between the species of interest and an outgroup species. In many species there is a significant positive correlation between D_S and P_S (Table 3) suggesting that part of the variation in diversity is due to variation in the mutation rate. However, if we test whether there is significant variation between loci taking into account the mutation rate, as estimated from the divergence between species, using either a χ^2 test of independence or the more conservative HKA test, then we find significant evidence in the majority of species, whether or not we assume free or no recombination within loci; 9 out of 10 loci for free recombination test and 6 out of 9 loci for the no recombination test (the HKA test could not be performed on *S.bicolor* due to the large number of genes in which the divergence was zero).

Correlates of N_e The variation in N_e across the genome is likely to be due to genetic hitch-hiking and background selection. Both processes are expected to be stronger in regions of the genome with low rates of recombination and a high density of sites subject to natural selection. To investigate which, if either of these factors is responsible for the variation in N_e , we investigated whether the variation in N_e was correlated to the rate of recombination and density of selected sites in four species for which this data was available: *D. melanogaster*, human, mouse and *A. thaliana*. We measured the density of selected sites as either the number of nucleotides in annotated exons (genic density), or the number of nucleotides in conserved regions (conserved site density), as annotated in the UCSC conservation track, in windows of size 50KB, 500KB and 5MB, where the window is centred on the gene from which the polymorphism data was taken (there is no conservation track for *A. thaliana*, so in this species we just investigated the density of genic sites). Results for the different window sizes were generally consistent, so we present the results from the 500KB window size. We estimated N_e as the synonymous diversity divided by synonymous divergence.

In *D. melanogaster* we find, as others have done, that our estimate of N_e is positively correlated to recombination rate (Spearman's correlation coefficient $r=0.45$, $P < 0.01$). It is however also positively correlated to the density of conserved sites ($r=0.24$, $P < 0.01$), which is unexpected, though not genic sites ($r=0.03$, $P = 0.65$). The positive correlation with conserved site density might be due to the positive correlation that exists between the density of conserved sites and the rate of recombination ($r=0.56$, $P < 0.01$), and indeed if we perform a multiple regression we find that the correlation between N_e and the density of conserved sites disappears ($P=0.74$), while the positive correlation between N_e and recombination rate remains ($P < 0.01$).

In humans we find, as others have done, that both diversity ($r=0.14$, $P=0.02$) and divergence ($r=0.18$, $P < 0.01$) are positively correlated to the rate of recombination (LERCHER and HURST, 2002; HELLMANN *et al.*, 2005), and there is, as a consequence, no correlation between estimates of N_e and the rate of recombination ($r=0.026$, $P=0.69$). N_e is significantly negatively correlated to the density of genic sites ($r=-0.19$, $P < 0.01$), but

not conserved sites ($r=-0.085$, $P = 0.17$). Using multiple regression does not alter this picture; N_e is only correlated to the density of genic sites.

In mouse we see no significant correlations between estimates of N_e and the rate of recombination ($r=0.054$, $P=0.72$), the density of genic ($r=0.089$, $P=0.53$) or conserved sites ($r=0.093$, $P=0.51$). This picture is unaffected by the use of multiple regression.

In *A. thaliana* we see a pattern like that in humans; both diversity ($r=0.10$, $P=0.04$) and to a lesser extent divergence ($r=0.064$, $P=0.11$) are positively correlated to recombination rate, and N_e is positively but not significantly correlated to recombination rate ($r=0.080$, $P=0.11$). N_e is significantly negatively correlated to genic density ($r=-0.11$, $P=0.02$). Unfortunately there is no data on conserved sites in this species.

Quantifying variation of N_e Since we find evidence for variation in N_e in many of our species we attempted to quantify the amount of variation using a hierarchical Bayesian model. We assume underlying distributions for N_e and μ (e.g. log-normal distributions) and estimate the shape parameters σ_{N_e} and σ_μ and hence the variances of these distributions; the mean of each distribution is constrained to be equal to one (see materials and methods). We investigate two different models: in the first we assume free recombination and in the second we assume no recombination within loci, but free recombination between loci. These two models are likely to set the upper and lower bounds on the true level of variation in N_e . Under the free recombination model all the variation in diversity is attributed to variation in N_e , variation in the mutation rate and sampling error. In the model with no recombination, variation in diversity may additionally be due to variation in the coalescent process. Hence, the free recombination model gives an upper estimate on the variation in N_e and the no recombination model gives a lower bound.

We applied our method to the polymorphism data from each of the 10 eukaryotic species to estimate the variation of N_e within each genome along with the variation in the mutation rate, σ_μ (Table 4). As expected in all cases the estimate of σ_{N_e} is larger when free recombination is assumed, but the estimates from the two models are highly correlated

($r=0.95$). The estimate of σ_μ is unaffected by the model of recombination assumed. We find evidence that the value of σ_{N_e} varies between species for both the free and no recombination models ($P = 2.5 \times 10^{-9}$ and $P = 4.2 \times 10^{-8}$ respectively). We find that the level of variation of N_e is the lowest for *Mus musculus* and highest for *Capsella rubella* for both recombination models. The estimates of σ_{N_e} and σ_μ were of similar magnitude for each taxon suggesting that overall variation in the mutation rate and variation in the effective population size contribute a similar amount to the variation in diversity.

The level of variation in N_e we estimate using our method is quite modest. For example, *C.rubella* has the highest estimate of σ_{N_e} , but under this distribution the genes in the 90th percentile have an N_e that is only 7.2-fold greater than those in the 10th percentile, i.e. 80% of genes have an effective population size within 7.2-fold of each other. Four species have estimates of σ_{N_e} of less than 0.6 meaning that the difference between the 90th and 10th percentile is less than 4-fold.

The estimated distribution appears to fit the data reasonably well (Figure 2). We would not expect the fit to be perfect, particularly at the lower end of the distribution, since this is where sampling error is a major issue; e.g. many genes have no polymorphism because of sampling error, not because they have an effective population size of zero. It is possible that assuming a log-normal distribution places some unwanted constraints on the estimation procedure; in particular the probability density tends to zero for low N_e . We therefore also fitted a gamma distribution to the data (Table S3); with this distribution the probability density does not necessarily decline to zero near the origin. However, the estimated distributions are very similar to those obtained assuming a log-normal distribution (Figures S4 and S5). The species which show low variation in N_e are also those which tend to show little evidence of variation in N_e , as judged by the χ^2 or HKA tests. This implies that failure to detect variation in N_e is largely because there is limited variation in N_e rather issues with statistical power.

Variation in the efficiency of selection Although we estimate the variation in the effective population size to be modest, it is of interest to investigate whether this translates

into significant differences in the efficiency of natural selection across the genome. To investigate this we tested whether there was a correlation between $\psi = P_n/(P_s + 1)$ and either θ_s or N_e for each locus in a manner which controls for the obvious non-independence of the two variables (see materials and methods). We remove the non-independence by splitting P_s into two independent parts and we use ψ because it reduces the bias inherent in the estimation of P_n/P_s ; furthermore it allows P_n/P_s to be calculated for all genes (PIGANEAU and EYRE-WALKER, 2009). This test is not very powerful since ψ has a large variance; furthermore it is statistically biased in a manner which tends to generate a positive correlation between ψ and θ_s or N_e . We therefore follow the approach suggested by (PIGANEAU and EYRE-WALKER, 2009) and grouped genes according to their θ or N_e value. The results are qualitatively similar for groupings of 4, 8 and 16 genes, so we present the results for groups of 4. There is a significant negative correlation between both θ_s and ψ and N_e and ψ in *A. lyrata* and *C. grandiflora*, and a marginally significant correlation between ψ and θ_s in *D. melanogaster*, although only the correlations in *C. grandiflora* are significant after correction for multiple tests; otherwise the correlations are generally weak and non-significant. However, overall we find significant evidence for a negative correlation between ψ and θ_s or N_e if we combine probabilities: between ψ and θ_s $P=0.043$ and between ψ and N_e $P=0.021$.

The relationship between ψ and N_e can potentially yield information about the distribution of fitness effects (DFE; LOEWE *et al.*, 2006; LOEWE and CHARLESWORTH, 2006; WOOLFIT, 2006; ELYASHIV *et al.*, 2010). If we assume that the DFE for non-synonymous mutations is a gamma distribution, and that synonymous mutations are neutral, then P_n/P_s is expected to be proportional to $N_e^{-\beta}$, where β is the shape parameter of the gamma distribution (WELCH *et al.*, 2008). Hence we can estimate β by considering the slope of the regression line between $\log(\psi)$ and $\log(N_e)$. Since the log of zero is undefined we grouped genes in groups of size n such that no group had a zero estimate of ψ or N_e . We attempted to estimate β in the species which individually showed a significant correlation between ψ and N_e . However, we could not perform the analysis of *A. lyrata* because the diversity is so low that it was impossible to define groups that did not have

zero values for both ψ and N_e . The estimates of β using this method are 0.41 (SE=0.15) in *C. grandiflora* and 0.23 (0.15) in *D. melanogaster*; these are similar to those obtained using an independent method that uses the site frequency spectrum (KEIGHTLEY and EYRE-WALKER, 2007): 0.27 (0.08) for *C. grandiflora* and 0.29 (0.07) for *D. melanogaster* (Table S2). This suggests that the gamma distribution is a reasonable approximation to the DFE, at least for mutations of weak effect.

Discussion

The effective population size (N_e) is one of the most important parameters in population genetics and evolutionary biology. It has been shown that N_e varies across the genome of *Drosophila melanogaster* and some plant species, and it is thought that it might vary across the human genome (HELLMANN *et al.*, 2005). Here we have shown that it varies in most species that we have considered. However, the variation in N_e is not consistently correlated to either the rate of recombination or the density of selected sites. This might in part be because the variation in N_e is quite limited; most genes in a genome have an N_e which is within a few fold of most other genes. Nevertheless the variation is sufficient to cause differences in the effectiveness of natural selection on segregating non-synonymous polymorphisms.

There are a number of factors which might have led us to over- or underestimate the variation in N_e . First, we have assumed that there is either free recombination or no recombination within loci to estimate the variation in the effective population size. This is unsatisfactory since we know that recombination is one of the factors which generates variation in the effective population size, at least in species like *Drosophila*, in which there is a correlation between diversity and the rate of recombination. Unfortunately it is not easy to get around this problem. However, as we have noted earlier, the estimate assuming free recombination should give an upper estimate on the amount of variation, because under this method all variation in the diversity is assumed to arise from sampling error and variation in the mutation rate and N_e . In reality, some of the variation between

genes will be a consequence of variation in the length of the genealogy in genes with little or no recombination.

Second, we have used the divergence between species as an estimate of the mutation rate, but if the mutation rate at a locus changes through time, for which there is evidence (AGUILETA *et al.*, 2006; HODGKINSON and EYRE-WALKER, 2011), then we will tend to overestimate the variation in N_e ; this is most easily seen by assuming there is variation in the mutation rate, but no variation in N_e ; if the mutation rate has changed through time then the divergence will not be a perfect measure of the recent mutation rate and there will appear to be variation in N_e .

Third, we have assumed that synonymous mutations are neutral, but there is evidence of selection in humans (IDA and AKASHI, 2000) and other species (DURET, 2002; POND and MUSE, 2005); although it is clear that selection has acted upon synonymous mutations in the past in *Drosophila melanogaster*, the evidence of selection currently acting is contradictory (AKASHI, 1996; MCVEAN and VIEIRA, 2001; ZENG and CHARLESWORTH, 2010) and biased gene conversion may be acting (GALTIER *et al.*, 2006; ZENG and CHARLESWORTH, 2010). Most of the other species we have analysed have not been investigated in any detail. We need to consider two models. In the first model, let us assume that there is no variation in N_e but that there is variation in the strength of selection on synonymous codons. Such a model would generate apparent variation in N_e with the genes subject to the strongest selection apparently having the highest N_e , because negative selection affects divergence to a greater extent than polymorphism (KIMURA, 1983). However, this would lead to the regions of the genome with the lowest diversity apparently having the highest effective population size. This is clearly not the case; if we split P_s into two independent samples, using a hypergeometric distribution, then we find a positive correlation between our estimate of N_e and P_s (Table 3). In the second model, let us imagine that there is variation in N_e and variation in the strength of selection on codon usage bias, but that they are uncorrelated to each other. In this case selection on codon usage bias will tend to generate an overestimate of the variation in N_e : as N_e increases selection becomes more effective, but this reduces the divergence more than the level of polymorphism, yielding a

higher apparent effective population size. So genes in regions of high N_e will tend to have an exaggerated N_e . There is also another effect that needs to be considered. We have estimated the level of synonymous divergence using the method of Goldman and Yang (GOLDMAN and YANG, 1994; YANG and NIELSEN, 1998), which assumes that codon bias is due to mutation bias; however, this method will tend to overestimate the synonymous substitution rate if codon bias is due to selection, because it will incorrectly infer that genes with high bias have a small number of synonymous sites, and hence a relatively large number of substitutions (BIERNE and EYRE-WALKER, 2003; YANG, 2006). As a consequence the divergence in high biased genes will be overestimated, but at the same time the mutation rate will tend to be underestimated because of the action of selection. These two factors may cancel each other out.

Fourth, we have only applied our method to protein coding sequences, so we are estimating the variation in the effective population size that applies to the proteome; there might be further variation in N_e in regions that are relatively devoid of protein coding sequences, such as heterochromatin. Whether this is important depends on whether there are functional sequences within these regions. We have also only considered genes on the autosomes and occasionally the homogametic sex-chromosome (14 loci in *H.sapiens*). We have not considered genes on the heterogametic sex chromosome, which often appear to have much lower effective population sizes. However, the heterogametic sex chromosome usually has very few genes (GRAVES, 2006).

Fifth, in estimating the variation in N_e we have assumed that there is either free recombination or no recombination and the population size has been stationary. Variation in population size can generate variation in diversity between loci, which may for example be mistaken for the signature of genetic hitch-hiking (TAJIMA, 1989; PLUZHNIKOV *et al.*, 2002). In principle we could take this into account by estimating a demographic model from the polymorphism data while simultaneously estimating the variation in N_e . This is difficult and is beyond the scope of the current work.

Finally, we have not taken into account ancestral polymorphism within our method. Ig-

noring ancestral polymorphism will lead us to underestimate the variation in N_e because loci with large N_e will tend to have higher divergences than loci with small N_e and this will appear as though these loci have higher mutation rates; variation in N_e will therefore be underestimated because the mutation rate has been overestimated. In principle it is possible to include ancestral polymorphism within the method, but we observe a lack of convergence, probably because the number of polymorphisms for each gene was so low. However, we have chosen datasets in which divergence is generally considerably larger than diversity; for example, we chose macaque as the outgroup to humans because variation in N_e does appear to generate variation in the divergence between human and chimpanzee (McVICKER *et al.*, 2009).

Despite finding variation in N_e in many of the species we tested, we find no consistent evidence that N_e is correlated to either the rate of recombination or the density of selected sites, the two factors which we would have expected variation in N_e to depend upon. This is probably in part due to the fact that we are using synonymous diversity; as such our estimates of diversity are subject to considerable error. The lack of a strong correlation between recombination rate and N_e may also be due to the fact that the genetic maps in *A. thaliana* and mouse are relatively crude. Furthermore, for our mouse species we are using an F2 genetic linkage map constructed from intercrosses between *M. m. domesticus* and *M. m. castaneus* to infer recombination rates for *Mus musculus castaneus*. In humans it has previously been shown that diversity over divergence is correlated positively to recombination rate (HELLMANN *et al.*, 2005) and that d_n/d_s is correlated to gene density (BULLAUGHEY *et al.*, 2008). In contrast to HELLMANN *et al.* (2005) we do not find a significant correlation between N_e and recombination rate, but they used long non-coding sequences to investigate diversity over divergence; their estimates were therefore subject to much less error than ours. It is surprising that there is a correlation between genic density but not conserved site density in humans. This might be due to the fact that there is approximately twice as much variation in genic density as conserved site density (coefficient of variation 0.79 versus 0.30). It might also be due to differences in the DFE between the two types of sites; background selection is most effective when the

strength of selection acting upon deleterious mutations is similar in magnitude to the rate of recombination (NORDBORG *et al.*, 1996).

In contrast genetic hitch-hiking depends upon the rate of advantageous mutation and sequences undergoing considerable adaptive evolution may not appear as conserved; the correlation between N_e and the density of genic sites may therefore suggest that hitch-hiking is more important in generating variation in N_e , than background selection. The lack of a correlation between N_e and the density of selected sites in *Drosophila*, once correlations to the rate of recombination have been taken into account may reflect the fact that the variation in N_e is generated by genetic hitch-hiking and a lot of adaptive evolution goes on outside coding sequences (ANDOLFATTO, 2005).

Across species we find evidence that variation in N_e leads to variation in the effectiveness of natural selection on non-synonymous mutations across the genome (Table 5). However, this is individually significant for just two genomes: *C. grandiflora* and *A. lyrata*. A lack of a correlation in other genomes may be due to the fact that we have little power to detect the correlation since i) some of the datasets are quite small, ii) there is limited variation in N_e and iii) in most of these species the DFE is very leptokurtic. The kurtosis of the DFE is such that changes in effective population size do not greatly change the proportion of mutations that are effectively neutral. It can be shown that under a gamma DFE the proportion of effectively neutral mutations is proportional to $N_e^{-\beta}$ (OHTA, 1977; KIMURA, 1979, 1983; WELCH *et al.*, 2008). Since β values are typically between 0.1 and 0.3 in most species (Table S2), changes in N_e tend to cause small changes in the proportion of effectively neutral mutations; for example a ten-fold increase in effective population size will reduce the proportion of effectively neutral mutations by only 37% if beta is 0.2. We find no evidence of a significant negative correlation between ψ and either θ_S or N_e in humans, in agreement with the work of (BULLAUGHEY *et al.*, 2008). They found no evidence that the ratio of the non-synonymous (d_N) to the synonymous (d_S) substitution rate between human, chimpanzee and macaque was correlated to the rate of recombination.

We find evidence that the amount of variation in N_e varies between species, however there are no obvious correlates of this variation. Both plants and animals have species with high and low levels of variation. Surprisingly we find no obvious effect of self-fertilization as suggested by previous studies (CUTTER and PAYSEUR, 2003; ROSELIUS *et al.*, 2005). *A.thaliana*, *C.rubella* and *B.stricta* are all self-fertile with selfing rates of around 0.95, 1 and 0.94 respectively (CHARLESWORTH and VEKEMANS, 2005; SONG *et al.*, 2006; FOXE *et al.*, 2009), whereas the closely related species *A. lyrata* and *C. grandiflora* are obligate outcrossing species. However, the variation in N_e seems to be relatively low for *C.grandiflora* and *B.stricta* and similar for the two *Arabidopsis* species. It also should be noted that the confidence intervals on the estimate of N_e in *C.rubella* are very large and a substantial amount of variation is still shared between *C.grandiflora* and *C.rubella* so these estimates are not independent. Moreover, the lack of an effect for self-compatibility in our estimates of N_e for *Arabidopsis* may be not surprising as self-compatibility might have been evolved relatively recently in *Arabidopsis* (BECHSGAARD *et al.*, 2006; TANG *et al.*, 2007). Furthermore, both *Arabidopsis* species have high sequence diversity in pericentromeric regions (BOREVITZ *et al.*, 2007; KAWABE *et al.*, 2008) which is not caused by varying mutation rates. Therefore this could be a major determinant of variation in N_e in those species and interfere with the effects of the breeding system.

Although the variation we observe in the effective population size appears to be modest, it does appear to influence both the level of neutral genetic diversity, and the effectiveness of selection. This potentially has important implications. If slightly deleterious mutations contribute substantially to phenotypic traits, then variation in the effective population size may affect where the genetic variation underlying fitness and other traits is distributed. For example, (ROCKMAN *et al.*, 2010) have recently shown that expression QTLs (eQTLs) tend to be present in regions of the *C.elegans* genome with the highest rates of recombination and lowest density of genes, where N_e is expected to be largest. However, population genetic theory also suggests that such weakly selected mutations are unlikely to contribute much to the overall genetic variance in fitness unless the proportion of mutations under such weak selection is large (EYRE-WALKER, 2010). Variation in the

effective population might also affect the rate of adaptive evolution, as appears to be the case in *Drosophila* (BETANCOURT and PRESGRAVES, 2002). Advantageous mutations can potentially come from three sources. They can be generated de novo, in which case we expect regions of the genome to adapt faster because the number of chromosomes an advantageous mutation can occur in is larger, and selection will be more effective on a greater proportion of the advantageous mutations. Advantageous mutations can also arise from standing genetic variation (PRITCHARD *et al.*, 2010; PRITCHARD and RIENZO, 2010). If these mutations were previously strongly deleterious, the genetic variation is not expected to depend upon N_e , unless the mutations are highly recessive. If, however, the advantageous mutations were previously neutral or weakly selected, regions of the genome with high N_e are expected to have more genetic variation and hence adapt more rapidly.

Acknowledgments

We are grateful to several anonymous referees for comments. TIG was financially supported by the John Maynard Smith studentship.

Literature Cited

- AGUILETA, G., J. P. BIELAWSKI, and Z. YANG, 2006 Evolutionary rate variation among vertebrate beta globin genes: implications for dating gene family duplication events. *Gene* **380**: 21–29.
- AKASHI, H., 1996 Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. *Genetics* **144**: 1297–1307.
- ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN, 1990 Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

- ANDOLFATTO, P., 2005 Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437**: 1149–1152.
- ARGUELLO, J. R., Y. ZHANG, T. KADO, C. FAN, R. ZHAO, *et al.*, 2010 Recombination yet inefficient selection along the *Drosophila melanogaster* subgroup's fourth chromosome. *Mol Biol Evol* **27**: 848–861.
- BACHTROG, D., and B. CHARLESWORTH, 2002 Reduced adaptation of a non-recombining neo-Y chromosome. *Nature* **416**: 323–326.
- BECHSGAARD, J. S., V. CASTRIC, D. CHARLESWORTH, X. VEKEMANS, and M. H. SCHIERUP, 2006 The transition to self-compatibility in *Arabidopsis thaliana* and evolution within S-haplotypes over 10 Myr. *Mol Biol Evol* **23**: 1741–1750.
- BEGUN, D. J., and C. F. AQUADRO, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**: 519–520.
- BERLIN, S., and H. ELLEGREN, 2006 Fast accumulation of nonsynonymous mutations on the female-specific W chromosome in birds. *J Mol Evol* **62**: 66–72.
- BETANCOURT, A. J., and D. C. PRESGRAVES, 2002 Linkage limits the power of natural selection in *Drosophila*. *Proc Natl Acad Sci U S A* **99**: 13616–13620.
- BETANCOURT, A. J., J. J. WELCH, and B. CHARLESWORTH, 2009 Reduced effectiveness of selection caused by a lack of recombination. *Curr Biol* **19**: 655–660.
- BIERNE, N., and A. EYRE-WALKER, 2003 The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics* **165**: 1587–1597.
- BOREVITZ, J. O., S. P. HAZEN, T. P. MICHAEL, G. P. MORRIS, I. R. BAXTER, *et al.*, 2007 Genome-wide patterns of single-feature polymorphism in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A* **104**: 12057–12062.

- BULLAUGHEY, K., M. PRZEWORSKI, and G. COOP, 2008 No effect of recombination on the efficacy of natural selection in primates. *Genome Res* **18**: 544–554.
- CHARLESWORTH, B., 2009 Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**: 195–205.
- CHARLESWORTH, B., M. T. MORGAN, and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–1303.
- CHARLESWORTH, D., and X. VEKEMANS, 2005 How and when did *Arabidopsis thaliana* become highly self-fertilising. *Bioessays* **27**: 472–476.
- CUTTER, A. D., and B. A. PAYSEUR, 2003 Selection at linked sites in the partial selfer *Caenorhabditis elegans*. *Mol Biol Evol* **20**: 665–673.
- DUMONT, B. L., M. A. WHITE, B. STEFFY, T. WILTSHIRE, and B. A. PAYSEUR, 2011 Extensive recombination rate variation in the house mouse species complex inferred from genetic linkage maps. *Genome Res* **21**: 114–125.
- DURET, L., 2002 Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev* **12**: 640–649.
- ELYASHIV, E., K. BULLAUGHEY, S. SATTATH, Y. RINOTT, M. PRZEWORSKI, *et al.*, 2010 Shifts in the intensity of purifying selection: an analysis of genome-wide polymorphism data from two closely related yeast species. *Genome Res* **20**: 1558–1573.
- EYRE-WALKER, A., 1996 Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? *Mol Biol Evol* **13**: 864–872.
- EYRE-WALKER, A., 2010 Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc Natl Acad Sci U S A* **107 Suppl 1**: 1752–1756.
- FILATOV, D. A., V. LAPORTE, C. VITTE, and D. CHARLESWORTH, 2001 DNA diversity in sex-linked and autosomal genes of the plant species *Silene latifolia* and *Silene dioica*. *Mol Biol Evol* **18**: 1442–1454.

- FOX, J. P., T. SLOTTE, E. A. STAHL, B. NEUFFER, H. HURKA, *et al.*, 2009 Recent speciation associated with the evolution of selfing in *Capsella*. *Proc Natl Acad Sci U S A* **106**: 5241–5245.
- FOX, J. P., V. UN NISA DAR, H. ZHENG, M. NORDBORG, B. S. GAUT, *et al.*, 2008 Selection on amino acid substitutions in *Arabidopsis*. *Mol Biol Evol* **25**: 1375–1383.
- GALTIER, N., E. BAZIN, and N. BIERNE, 2006 GC-biased segregation of noncoding polymorphisms in *Drosophila*. *Genetics* **172**: 221–228.
- GARRIGAN, D., and M. F. HAMMER, 2006 Reconstructing human origins in the genomic era. *Nat Rev Genet* **7**: 669–680.
- GOLDMAN, N., and Z. YANG, 1994 A codon-based model of nucleotide substitution for protein-coding dna sequences. *Mol Biol Evol* **11**: 725–736.
- GOSSMANN, T. I., B.-H. SONG, A. J. WINDSOR, T. MITCHELL-OLDS, C. J. DIXON, *et al.*, 2010 Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol* **27**: 1822–1832.
- GRAVES, J. A. M., 2006 Sex chromosome specialization and degeneration in mammals. *Cell* **124**: 901–914.
- GUO, Y.-L., J. S. BECHSGAARD, T. SLOTTE, B. NEUFFER, M. LASCoux, *et al.*, 2009 Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. *Proc Natl Acad Sci U S A* **106**: 5246–5251.
- HALLIGAN, D. L., F. OLIVER, A. EYRE-WALKER, B. HARR, and P. D. KEIGHTLEY, 2010 Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genet* **6**: e1000825.
- HAMBLIN, M. T., A. M. CASA, H. SUN, S. C. MURRAY, A. H. PATERSON, *et al.*, 2006 Challenges of detecting directional selection after a bottleneck: lessons from *Sorghum bicolor*. *Genetics* **173**: 953–964.

- HASTINGS, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97–109.
- HELLBORG, L., and H. ELLEGREN, 2004 Low levels of nucleotide diversity in mammalian Y chromosomes. *Mol Biol Evol* **21**: 158–163.
- HELLMANN, I., I. EBERSBERGER, S. E. PTAK, S. PÄÄBO, and M. PRZEWORSKI, 2003 A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet* **72**: 1527–1535.
- HELLMANN, I., K. PRÜFER, H. JI, M. C. ZODY, S. PÄÄBO, *et al.*, 2005 Why do human diversity levels vary at a megabase scale? *Genome Res* **15**: 1222–1231.
- HEY, J., and R. M. KLIMAN, 2002 Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. *Genetics* **160**: 595–608.
- HODGKINSON, A., and A. EYRE-WALKER, 2011 Variation in the mutation rate across the mammalian genome. *Nat Rev Genet* : in press.
- HUDSON, R. R., M. KREITMAN, and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153–159.
- IIDA, K., and H. AKASHI, 2000 A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene* **261**: 93–105.
- INGVARSSON, P. K., 2004 Population subdivision and the Hudson-Kreitman-Aguade test: testing for deviations from the neutral model in organelle genomes. *Genet Res* **83**: 31–39.
- INNAN, H., 2006 Modified Hudson-Kreitman-Aguade test and two-dimensional evaluation of neutrality tests. *Genetics* **173**: 1725–1733.
- KAWABE, A., A. FORREST, S. I. WRIGHT, and D. CHARLESWORTH, 2008 High DNA sequence diversity in pericentromeric genes of the plant *Arabidopsis lyrata*. *Genetics* **179**: 985–995.

- KEIGHTLEY, P. D., and A. EYRE-WALKER, 2007 Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**: 2251–2261.
- KIMURA, M., 1979 Model of effectively neutral mutations in which selective constraint is incorporated. *Proc Natl Acad Sci U S A* **76**: 3440–3444.
- KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press.
- KLIMAN, R. M., and J. HEY, 2003 Hill-Robertson interference in *Drosophila melanogaster*: reply to Marais, Mouchiroud and Duret. *Genet Res* **81**: 89–90.
- KONG, A., D. F. GUDBJARTSSON, J. SAINZ, G. M. JONSDOTTIR, S. A. GUDJONSSON, *et al.*, 2002 A high-resolution recombination map of the human genome. *Nat Genet* **31**: 241–247.
- LARRACUENTE, A. M., T. B. SACKTON, A. J. GREENBERG, A. WONG, N. D. SINGH, *et al.*, 2008 Evolution of protein-coding genes in *Drosophila*. *Trends Genet* **24**: 114–123.
- LERCHER, M. J., and L. D. HURST, 2002 Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* **18**: 337–340.
- LITI, G., D. M. CARTER, A. M. MOSES, J. WARRINGER, L. PARTS, *et al.*, 2009 Population genomics of domestic and wild yeasts. *Nature* **458**: 337–341.
- LOEWE, L., and B. CHARLESWORTH, 2006 Inferring the distribution of mutational effects on fitness in *Drosophila*. *Biol Lett* **2**: 426–430.
- LOEWE, L., B. CHARLESWORTH, C. BARTOLOMÉ, and V. NÖEL, 2006 Estimating selection on nonsynonymous mutations. *Genetics* **172**: 1079–1092.
- MACHADO, C. A., R. M. KLIMAN, J. A. MARKERT, and J. HEY, 2002 Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. *Mol Biol Evol* **19**: 472–488.

- MARAIS, G., D. MOUCHIROUD, and L. DURET, 2003 Neutral effect of recombination on base composition in *Drosophila*. *Genet Res* **81**: 79–87.
- MCVEAN, G. A., and J. VIEIRA, 2001 Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. *Genetics* **157**: 245–257.
- MCVICKER, G., D. GORDON, C. DAVIS, and P. GREEN, 2009 Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet* **5**: e1000471.
- MONTELL, H., A. K. FRIDOLFSSON, and H. ELLEGREN, 2001 Contrasting levels of nucleotide diversity on the avian Z and W sex chromosomes. *Mol Biol Evol* **18**: 2010–2016.
- MORIYAMA, E. N., and J. R. POWELL, 1996 Intraspecific nuclear DNA variation in *Drosophila*. *Mol Biol Evol* **13**: 261–277.
- NORDBORG, M., B. CHARLESWORTH, and D. CHARLESWORTH, 1996 The effect of recombination on background selection. *Genet Res* **67**: 159–174.
- NORDBORG, M., T. T. HU, Y. ISHINO, J. JHAVERI, C. TOOMAJIAN, *et al.*, 2005 The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol* **3**: e196.
- OHTA, T., 1977 Extension to the nearly neutral random drift hypothesis. Kimura M (ed) *Evolution and polymorphism*, National Institute of Genetics, Mishima : 148–167.
- PIGANEAU, G., and A. EYRE-WALKER, 2009 Evidence for variation in the effective population size of animal mitochondrial DNA. *PLoS One* **4**: e4396.
- PLUZHNIKOV, A., A. D. RIENZO, and R. R. HUDSON, 2002 Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics* **161**: 1209–1218.
- POND, S. K., and S. V. MUSE, 2005 Site-to-site variation of synonymous substitution rates. *Mol Biol Evol* **22**: 2375–2385.

- PRESGRAVES, D. C., 2005 Recombination enhances protein adaptation in *Drosophila melanogaster*. *Curr Biol* **15**: 1651–1656.
- PRITCHARD, J. K., J. K. PICKRELL, and G. COOP, 2010 The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol* **20**: R208–R215.
- PRITCHARD, J. K., and A. D. RIENZO, 2010 Adaptation - not by sweeps alone. *Nat Rev Genet* **11**: 665–667.
- ROCKMAN, M. V., S. S. SKROVANEK, and L. KRUGLYAK, 2010 Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. *Science* **330**: 372–376.
- ROSELIUS, K., W. STEPHAN, and T. STÄDLER, 2005 The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics* **171**: 753–763.
- ROSS-IBARRA, J., S. I. WRIGHT, J. P. FOXE, A. KAWABE, L. DEROSE-WILSON, *et al.*, 2008 Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE* **3**: e2411.
- SCHMID, K. J., S. RAMOS-ONSINS, H. RINGYS-BECKSTEIN, B. WEISSHAAR, and T. MITCHELL-OLDS, 2005 A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* **169**: 1601–1615.
- SHAPIRO, J. A., W. HUANG, C. ZHANG, M. J. HUBISZ, J. LU, *et al.*, 2007 Adaptive genic evolution in the *Drosophila* genomes. *Proc Natl Acad Sci U S A* **104**: 2271–2276.
- SIEPEL, A., G. BEJERANO, J. S. PEDERSEN, A. S. HINRICHS, M. HOU, *et al.*, 2005 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.
- SINGER, T., Y. FAN, H.-S. CHANG, T. ZHU, S. P. HAZEN, *et al.*, 2006 A high-resolution map of *Arabidopsis* recombinant inbred lines by whole-genome exon array hybridization. *PLoS Genet* **2**: e144.

- SLOTTE, T., J. P. FOXE, K. M. HAZZOURI, and S. I. WRIGHT, 2010 Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. *Mol Biol Evol* **27**: 1813–1821.
- SMITH, J. M., and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. *Genet Res* **23**: 23–35.
- SONG, B.-H., M. J. CLAUSS, A. PEPPER, and T. MITCHELL-OLDS, 2006 Geographic patterns of microsatellite variation in *Boechera stricta*, a close relative of *Arabidopsis*. *Mol Ecol* **15**: 357–369.
- SONG, B.-H., A. J. WINDSOR, K. J. SCHMID, S. RAMOS-ONSINS, M. E. SCHRANZ, *et al.*, 2009 Multilocus patterns of nucleotide diversity, population structure and linkage disequilibrium in *Boechera stricta*, a wild relative of *Arabidopsis*. *Genetics* **181**: 1021–1033.
- STEPHAN, W., and H. LI, 2007 The recent demographic and adaptive history of *Drosophila melanogaster*. *Heredity* **98**: 65–68.
- STOLETZKI, N., and A. EYRE-WALKER, 2011 Estimation of the neutrality index. *Mol Biol Evol* **28**: 63–70.
- SUNDSTRÖM, H., M. T. WEBSTER, and H. ELLEGREN, 2004 Reduced variation on the chicken Z chromosome. *Genetics* **167**: 377–385.
- TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- TANG, C., C. TOOMAJIAN, S. SHERMAN-BROYLES, V. PLAGNOL, Y.-L. GUO, *et al.*, 2007 The evolution of selfing in *Arabidopsis thaliana*. *Science* **317**: 1070–1072.
- TENAILLON, M. I., J. U'REN, O. TENAILLON, and B. S. GAUT, 2004 Selection versus demography: a multilocus investigation of the domestication process in maize. *Mol Biol Evol* **21**: 1214–1225.

- THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673–4680.
- TSAI, I. J., D. BENSASSON, A. BURT, and V. KOUFOPANOU, 2008 Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. *Proc Natl Acad Sci U S A* **105**: 4957–4962.
- WELCH, J. J., A. EYRE-WALKER, and D. WAXMAN, 2008 Divergence and polymorphism under the nearly neutral theory of molecular evolution. *J Mol Evol* **67**: 418–426.
- WHITLOCK, M. C., 2005 Combining probability from independent tests: the weighted Z-method is superior to Fisher’s approach. *J Evol Biol* **18**: 1368–1373.
- WOOLFIT, M. R. Q., 2006 *Effective population size and its effects on molecular evolution*. University of Sussex School of Life Sciences Department of Biology.
- WRIGHT, S. I., and B. CHARLESWORTH, 2004 The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. *Genetics* **168**: 1071–1076.
- WYCKOFF, G. J., J. LI, and C.-I. WU, 2002 Molecular evolution of functional genes on the mammalian Y chromosome. *Mol Biol Evol* **19**: 1633–1636.
- YANG, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**: 555–556.
- YANG, Z., 2006 *Computational Molecular Evolution*. Oxford University Press, USA.
- YANG, Z., and R. NIELSEN, 1998 Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol* **46**: 409–418.
- ZENG, K., and B. CHARLESWORTH, 2010 Studying patterns of recent evolution at synonymous sites and intronic sites in *Drosophila melanogaster*. *J Mol Evol* **70**: 116–128.
- ZHANG, J., D. M. WEBB, and O. PODLAHA, 2002 Accelerated protein evolution and origins of human-specific features: *Foxp2* as an example. *Genetics* **162**: 1825–1835.

Figure Legends

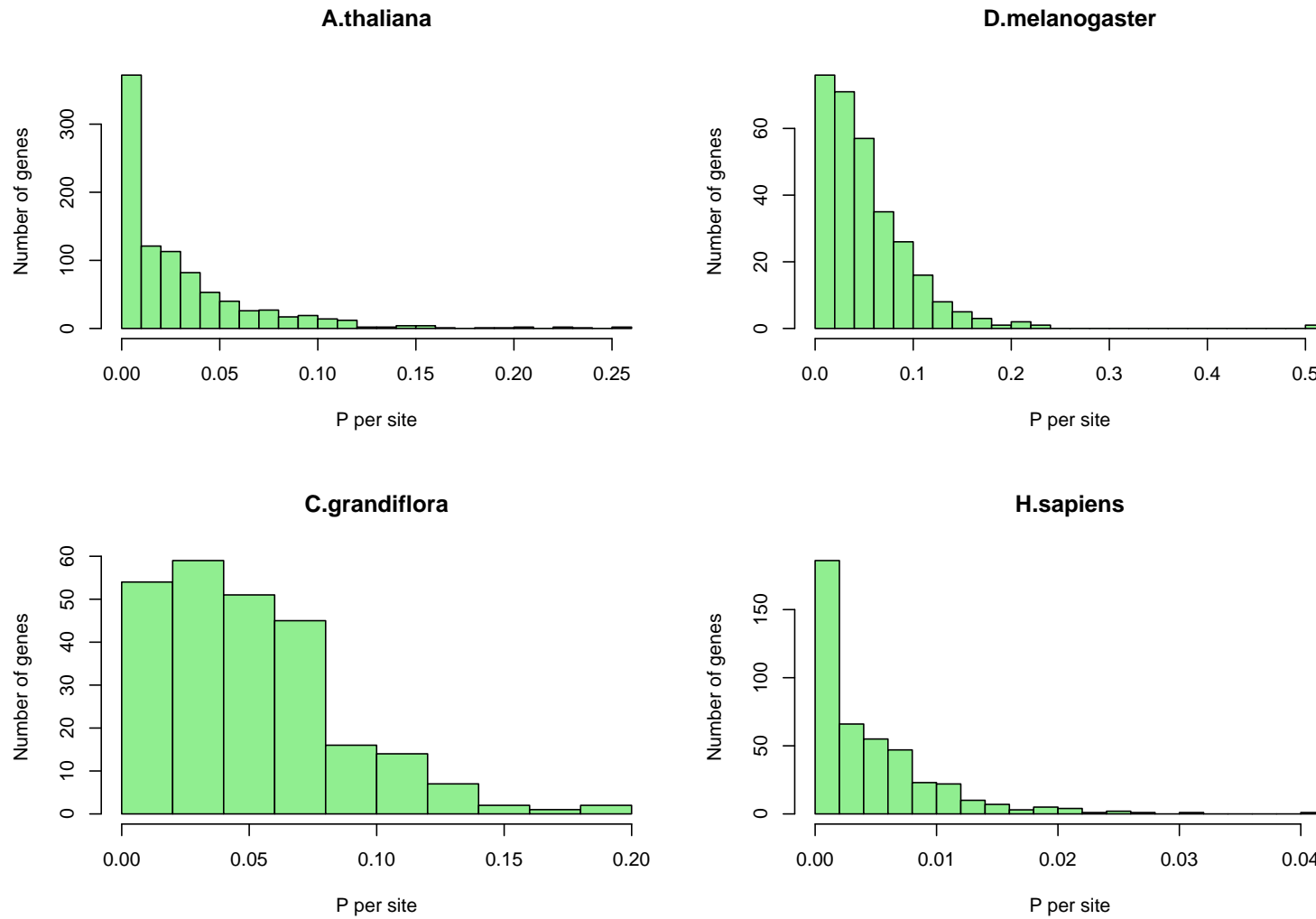


Figure 1: Distribution of the number of polymorphisms per site across genes for four species.

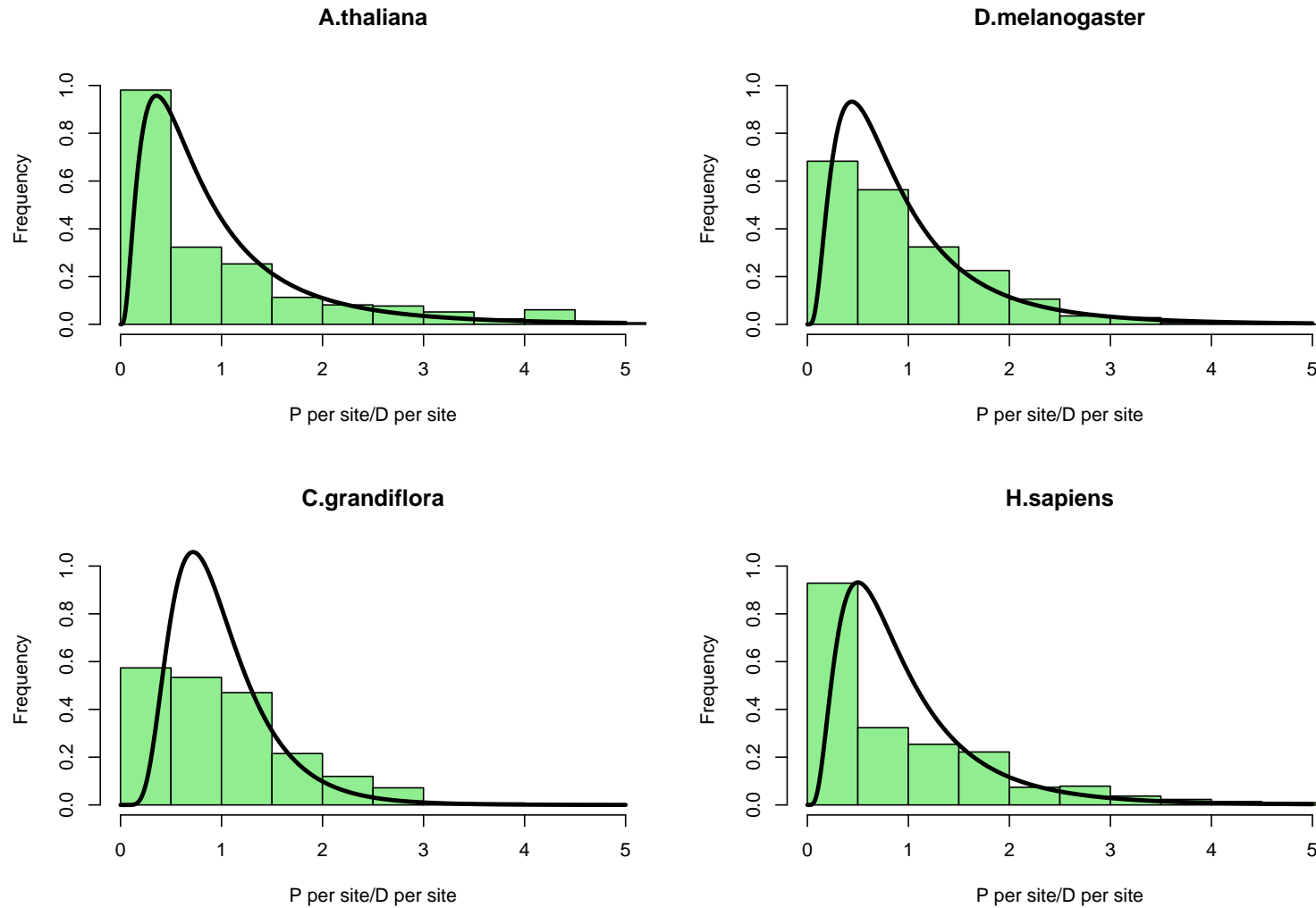


Figure 2: Distribution of the per site polymorphism/divergence ratio across genes for four species and corresponding distributions of N_e (solid line) estimated by the Hierarchical Bayesian analysis assuming a log-normal distribution.

Tables

Table 1: Summary of data sets used for the analyses

Species	Outgroup	Loci	Sites	Alleles	θ_s	d_s	Dataset
<i>Drosophila melanogaster</i>	<i>D.simulans</i>	302	40920	8	0.019	0.13	SHAPIRO <i>et al.</i> (2007)
<i>Homo sapiens</i>	<i>Macaca mulatta</i>	434	170441	32	0.001	0.08	EGP/PGA ¹
<i>Mus musculus</i>	<i>Rattus norvegicus</i>	66	5127	20	0.010	0.21	HALLIGAN <i>et al.</i> (2010)
<i>Arabidopsis thaliana</i>	<i>A.lyrata</i>	918	64927	24	0.008	0.14	NORDBORG <i>et al.</i> (2005)
<i>Capsella grandiflora</i>	<i>Neslia paniculata</i>	251	31273	8	0.019	0.16	SLOTTE <i>et al.</i> (2010)
<i>Sorghum bicolor</i>	<i>S.propinquum</i>	134	6799	14	0.004	0.02	HAMBLIN <i>et al.</i> (2006)
<i>Boechera stricta</i>	<i>A.thaliana</i>	129	10048	40	0.003	0.21	SONG <i>et al.</i> (2009); GOSSMANN <i>et al.</i> (2010)
<i>Arabidopsis lyrata</i>	<i>A.thaliana</i>	66	5260	24	0.018	0.15	ROSS-IBARRA <i>et al.</i> (2008); FOXE <i>et al.</i> (2008)
<i>Capsella rubella</i>	<i>A.thaliana</i>	49	5014	16	0.004	0.29	FOXSE <i>et al.</i> (2009); GUO <i>et al.</i> (2009)
<i>S.paradoxus</i>	<i>S.cerevisiae</i>	94	28019	8	0.002	0.36	TSAI <i>et al.</i> (2008)

¹ EGP: <http://egp.gs.washington.edu> and PGA: <http://pga.gs.washington.edu> August 2010

Number of synonymous sites (Sites) and nucleotide diversity (θ_s) are from the polymorphism data. Average divergence between the species pairs at silent sites (d_s).

Table 2: Results of the χ^2 tests of independence and HKA tests

Species	Diversity		Diversity and Divergence	
	P-value (χ^2)	P-value (HKA)	P-value (χ^2)	P-value (HKA)
<i>D.melanogaster</i>	$< 1 \times 10^{-3}$	0.015	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$
<i>H.sapiens</i>	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$
<i>M.musculus</i>	$< 1 \times 10^{-3}$	0.432*	0.066*	0.429*
<i>A.thaliana</i>	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$
<i>C.grandiflora</i>	$< 1 \times 10^{-3}$	0.462*	$< 1 \times 10^{-3}$	0.565*
<i>S.bicolor</i>	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$	5.3×10^{-3}	n.a.
<i>B.stricta</i>	$< 1 \times 10^{-3}$	0.434*	6×10^{-3}	0.01
<i>A.lyrata</i>	$< 1 \times 10^{-3}$	5.4×10^{-3}	$< 1 \times 10^{-3}$	2.3×10^{-3}
<i>C.rubella</i>	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$
<i>S.paradoxus</i>	1.9×10^{-3}	0.94*	$< 1 \times 10^{-3}$	0.35*

* not significant (P>0.05)

Results of the χ^2 tests of independence and HKA tests for diversity and diversity/divergence data. For details see material and methods. P-values are given for each species.

Table 3: Results of correlates of P_s

Species	P_s vs D_s		P_s vs N_e	
	ρ	P-value	ρ	P-value
<i>D.melanogaster</i>	0.18	3.82×10^{-03}	0.46	1.87×10^{-17}
<i>H.sapiens</i>	0.29	1.62×10^{-06}	0.38	3.02×10^{-16}
<i>M.musculus</i>	0.38	5.98×10^{-03}	0.33	3.55×10^{-03}
<i>A.thaliana</i>	0.16	1.13×10^{-04}	0.44	5.94×10^{-45}
<i>C.grandiflora</i>	0.35	3.00×10^{-08}	0.52	2.53×10^{-19}
<i>S.bicolor</i>	0.54	3.17×10^{-03}	0.40	7.31×10^{-07}
<i>B.stricta</i>	0.10*	4.02×10^{-01}	0.14*	6.18×10^{-02}
<i>A.lyrata</i>	0.22*	1.03×10^{-01}	0.62	1.21×10^{-08}
<i>C.rubella</i>	0.12*	6.34×10^{-01}	0.65	2.35×10^{-07}
<i>S.paradoxus</i>	0.04*	7.91×10^{-01}	0.42	1.01×10^{-05}

* not significant (P>0.05)

Results of the Spearman's rank correlates of P_s . The non-independence of P_s and N_e is taken into account by splitting the dataset into independent halves (see materials and methods). Correlation coefficients (ρ) and P-values are given for each species.

Table 4: Estimates of the variation of N_e in 10 eukaryotic species.

Species	Free recombination		No recombination	
	σ_μ (Std)	σ_{N_e} (Std)	σ_μ (Std)	σ_{N_e} (Std)
<i>D.melanogaster</i>	0.370 (0.024)	0.743 (0.048)	0.372 (0.024)	0.516 (0.072)
<i>H.sapiens</i>	0.522 (0.021)	0.682 (0.07)	0.52 (0.02)	0.578 (0.11)
<i>M.musculus</i>	0.369 (0.045)	0.35 (0.119)	0.372 (0.045)	0.247 (0.15)
<i>A.thaliana</i>	0.419 (0.015)	0.83 (0.04)	0.423 (0.015)	0.809 (0.065)
<i>C.grandiflora</i>	0.355(0.021)	0.475 (0.043)	0.351 (0.021)	0.165 (0.067)
<i>S.bicolor</i>	0.689 (0.092)	0.903 (0.263)	0.710 (0.095)	0.675 (0.292)
<i>B.stricta</i>	0.441 (0.039)	0.503 (0.174)	0.443 (0.0379)	0.411 (0.178)
<i>A.lyrata</i>	0.276 (0.053)	0.729 (0.119)	0.278 (0.054)	0.651 (0.139)
<i>C.rubella</i>	0.263 (0.042)	1.191 (0.21)	0.258 (0.043)	1.126 (0.243)
<i>S.paradoxus</i>	0.23 (0.023)	0.566 (0.208)	0.23 (0.0218)	0.387 (0.131)

Estimates of the variation of N_e in 10 eukaryotic species. Results are for an underlying Log-Normal distribution for N_e and μ assuming either free recombination or no recombination (see materials and methods). For each dataset the mean shape parameters σ_{N_e} and σ_μ and in parentheses their standard deviations (Std) obtained from the posterior distribution are given.

Table 5: The correlation of $P_n/(P_s+1)=\psi$ and θ_s and N_e respectively in 10 eukaryotic species.

Species	ψ vs. θ_s (groups of 4)			ψ vs. N_e (groups of 4)		
	n	ρ	P-value	n	ρ	P-value
<i>D.melanogaster</i>	77	-0.172	0.067	77	-0.1	0.194
<i>H.sapiens</i>	110	-0.068	0.239	110	0.016	0.564
<i>M.musculus</i>	18	-0.253	0.155	18	-0.261	0.147
<i>A.thaliana</i>	231	0.051	0.781	231	0.055	0.799
<i>C.grandiflora</i>	64	-0.357	0.002	64	-0.483	2.673×10^{-5}
<i>S.bicolor</i>	35	0.093	0.702	35	0.001	0.504
<i>B.stricta</i>	33	0.164	0.818	33	-0.168	0.175
<i>A.lyrata</i>	18	-0.477	0.022	18	-0.507	0.016
<i>C.rubella</i>	13	0.451	0.939	13	0.491	0.955
<i>S.paradoxus</i>	25	-0.219	0.146	25	-0.019	0.462
combined (Z-method)			0.043			0.021

The non-independence of ψ and θ_s is taken into account by splitting the dataset into independent halves (see materials and methods). Correlation coefficients (ρ) and P-values (one-tailed) are given for each species.