Population Genomics of *Daphnia pulex*

Michael Lynch,*[,1] Ryan Gutenkunst,† Matthew Ackerman,* Ken Spitze,* Zhiqiang Ye,* Takahiro Maruki,* and Zhiyuan Jia*

*Department of Biology, Indiana University, Bloomington, Indiana 47405 and †Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona 85721

ABSTRACT Using data from 83 isolates from a single population, the population genomics of the microcrustacean *Daphnia pulex* are described and compared to current knowledge for the only other well-studied invertebrate, *Drosophila melanogaster*. These two species are quite similar with respect to effective population sizes and mutation rates, although some features of recombination appear to be different, with linkage disequilibrium being elevated at short ($<$ 100 bp) distances in *D. melanogaster* and at long distances in *D. pulex*. The study population adheres closely to the expectations under Hardy–Weinberg equilibrium, and reflects a past population history of no more than a twofold range of variation in effective population size. Fourfold redundant silent sites and a restricted region of intronic sites appear to evolve in a nearly neutral fashion, providing a powerful tool for population genetic analyses. Amino acid replacement sites are predominantly under strong purifying selection, as are a large fraction of sites in UTRs and intergenic regions, but the majority of SNPs at such sites that rise to frequencies >0.05 appear to evolve in a nearly neutral fashion. All forms of genomic sites (including replacement sites within codons, and intergenic and UTR regions) appear to be experiencing an $\sim 2\times$ higher level of selection scaled to the power of drift in *D. melanogaster*, but this may in part be a consequence of recent demographic changes. These results establish *D. pulex* as an excellent system for future work on the evolutionary genomics of natural populations.

KEYWORDS *Daphnia*; genetic variation; linkage disequilibrium; population genomics; population size

BY the 1950s, population geneticists were endowedwith a substantial body of theory suggesting what patterns of molecular variation ought to look like, conditional on the underlying forces of selection, mutation, recombination, and random genetic drift. Although the earliest glimpses of population-level molecular variation began to emerge in the 1960s with the advent of allozyme electrophoresis for separating protein variants on the basis of electrical charge, the picture was murky owing to the possibility of selection on amino acid differences and uncertainties about the DNA-level underpinnings (Ayala 1976). Shortly thereafter, crude methods for inferring variation at the nucleotide level (*e.g.*, analysis of endonuclease restriction sites) were developed (Nei and Koehn 1983), and this was soon followed by the emergence of the polymerase chain reaction (PCR) method for amplifying short stretches of DNA for complete sequencing. Today, high-throughput sequencing provides a path toward characterizing the entire genomes of multiple individuals, removing most of the potential biases that existed with earlier techniques and opening up the field of empirical population genetics to studies of all organisms. Large surveys of population-level variation have begun to emerge for humans and several of the model systems in molecular, cell, and developmental biology (*e.g.*, *Drosophila*, *Arabidopsis*, *Caenorhabditis*, *Schizosaccharomyces*, and *Saccharomyces*).

Here, we extend this scattered phylogenetic coverage to a long-standing model organism in ecological, physiological, and ecotoxicological research, the aquatic microcrustacean *Daphnia pulex*. This species occupies hundreds of thousands of ponds and lakes throughout the northern hemisphere, and has been subject to substantial study at both the molecular marker and quantitative genetic levels (Lynch and Spitze 1994; Pfrender *et al.* 2000). A genome sequence is available for the species (Colbourne *et al.* 2011), as is a genetic map based on single-sperm sequencing (Xu *et al.* 2015). Knockdown of gene expression by RNA interference is possible (Kato *et al.* 2011a, 2012; Schumpert *et al.* 2015a), and methods for targeted gene mutagenesis are also available (Nakanishi *et al.* 2014; Naitou *et al.* 2015). As a consequence

Copyright © 2017 by the Genetics Society of America
doi: https://doi.org/10.1534/genetics.116.190611
Manuscript received April 19, 2016; accepted for publication November 16, 2016; published Early Online December 5, 2016.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.190611/-/DC1.
[1]Corresponding author: Department of Biology, 1001 E. 3rd St., Indiana University, Bloomington, IN 47405. E-mail: milynch@indiana.edu

of the availability of these tools and the life history features of *Da. pulex* (below), the species has proven useful for studies in diverse areas related to molecular and cellular genetics, *e.g.*, recombination and sex (Omilian *et al.* 2006; Eads *et al.* 2012), mutation (Keith *et al.* 2016), intron evolution (W. Li *et al.* 2009), aging (Schumpert *et al.* 2014, 2015b), epigenetics (Harris *et al.* 2012; LeBlanc and Medlock 2015), sex determination (Toyota *et al.* 2015), and development (Shiga *et al.* 2002; Kato *et al.* 2011b; Mahato *et al.* 2014).

Most *Daphnia*, including all individuals in the population in this study, reproduce by cyclical parthenogenesis. During most of the growing season, females reproduce by a form of ameiotic parthenogenesis (Hiruta *et al.* 2010; Hiruta and Tochinai 2012), which ensures that offspring are genomically identical to their mothers, except in rare events of mutation and mitotic gene conversion (Omilian *et al.* 2006). At spring and summer temperatures, the age at first reproduction is 7–10 days, and females are capable of producing consecutive clutches every 2–3 days for several weeks. When conditions deteriorate, mothers produce sons via environmental sex determination, and convert to the production of haploid eggs, which upon fertilization lead to diapausing embryos contained within a dessication-resistant structure called an ephippium. When deposited at the bottom of permanent lakes, these resting eggs can survive for several hundred years and, after subsequent hatching in the laboratory, provide a powerful resource for studying evolutionary processes on the timescale of thousands of generations (Frisch *et al.* 2014; Möst *et al.* 2015).

In an effort to establish *Da. pulex* as a model system for ecological and evolutionary genomics, we have embarked on a project to sequence the genomes of ~ 100 clonal isolates from each of ~ 30 populations throughout the geographic range of the species. Each population resides in a pond that dries up annually, enforcing recurrent episodes of resting-egg production in the late spring/early summer followed by hatching in the late winter/early spring in the following year. Here, we provide a detailed evaluation of the population genomic structure from the first study population, where possible drawing comparisons with prior results for the fruit fly *Drosophila melanogaster*, the only other arthropod with a well-documented population genomic architecture. The results show that these two species, one aquatic with well-defined population boundaries and the other terrestrial with fluid geographic structure, have remarkably similar population genetic features but also some significant differences.

Materials and Methods

Sample collection, DNA preparation, and sequencing

Individuals were collected from Kicka Pond (KAP), located near Danville, IL (latitude 40.1224; longitude −87.7366) in the spring of 2013. To maximize the likelihood that each isolate would represent a hatchling from a unique resting egg, adults were collected from the water column before substantial reproduction had occurred. Such hatchlings would have resulted from a mating in the immediately preceding year, or possibly earlier, as resting eggs can survive multiple years under certain conditions (although long-term survival seems unlikely in highly oxidized surface soil). After expanding individual isolates clonally in the lab, DNA was extracted from ~ 100 individuals of each genotype (starved overnight in food-free medium) using a cetyltrimethylammonium bromide method (Doyle and Doyle 1987). DNA fragments in the libraries from each clone were marked with unique oligomer barcodes, and paired-end reads of length 100 or 150 bp (average insert size of 344 bp) were obtained by sequencing on an Illumina HiSequation 2500 platform.

Read mapping and data filtering

All sequences were binned according to their clone-specific barcodes, and then mapped to the reference genome (PA42; Z. Ye, S. Xu, K. Spitze, J. Asselman, X. Jiang, M. E. Pfrender, and M. Lynch, unpublished results), which is a high-coverage assembly derived from multifaceted data for a *Da. pulex* clone obtained from a nearby location. We used Novoalign (http://www.novocraft.com) using option "r None," excluding reads that mapped to more than one location on the reference genome, to align the fastq files to the PA42 reference assembly; Samtools (H. Li *et al.* 2009) to convert sam to bam files; GATK (McKenna *et al.* 2010) to remove PCR duplicates and perform local realignment around indels; clipOverlap (http://genome.sph.umich.edu/wiki/BamUtil: clipOverlap) to remove overlapping reads; and Samtools to generate Mpileup files. From the Mpileup files, the data were reduced to quartet sets of reads (numbers of observed As, Cs, Gs, and Ts at each site) for each clonal isolate using Mapgd, a developing package of computational methods for the analysis of population genomic data (M. S. Ackerman, P. Johri, K. Spitze, S. Xu, T. Doak, K. Young and M. Lynch, unpublished results). These quartet files were used in all subsequent analyses associated with the estimation of allele- and genotype-frequencies and linkage disequilibrium (LD).

In an initial screen of the data, preliminary maximum likelihood (ML) estimates of the genotype frequencies and the composite error rates (resulting from all sources of error, including amplification and sequencing) were obtained for each site, using the method of Maruki and Lynch (2015) implemented in Mapgd. We then performed a more detailed examination of each site to evaluate the goodness of fit of the clone-specific read data to the expectations under the ML model (which assumes biparental chromosome sampling and a random error distribution); among other things, this procedure was implemented to guard against clonal contamination during the period of laboratory maintenance and/or sample preparation. We also applied a ML method for estimating pairwise relatedness (M. S. Ackerman, P. Johri, K. Spitze, S. Xu, T. Doak, K. Young and M. Lynch, unpublished results) to guard against the inadvertent inclusion of clone mates in the original clone collection. Of the original 96 clones subjected to sequencing, nine were eliminated owing to having inadequate coverage, and four were deemed to

yield distributions of read data significantly divergent from model expectations (possibly due to contamination). This left a pool of 83 individuals for further analysis.

To minimize potential issues with mismappings of reads to paralogous or repetitive regions and to avoid sites subject to high error rates, based on the overall distribution of coverage across the genome, we further restricted all analyses to sites having 800–2400 × coverage for the entire population (∼ 50 and 150% of the modal population-level coverage across the genome) and error-rate estimates < 0.01. As discussed below, these error rates are estimated directly from the ML method used to estimate site-specific features. On rare occasions, a site will behave aberrantly within an individual (*e.g.*, being superamplified in the DNA sample). To avoid any downstream analysis problems associated with such issues, for each site that survived the criteria just noted, we further applied a ML procedure to detect individuals harboring read quartets deviating significantly from the preliminary-model expectations (the ML estimates of genotype frequencies and error rate at the site; M. S. Ackerman, P. Johri, K. Spitze, S. Xu, T. Doak, K. Young and M. Lynch, unpublished results). Provided no more than three individuals deviated from the model for the site, these were purged from the site-specific data, and the ML estimates of the genotype frequencies were redetermined; if four or more individuals were deemed to have inadequate fits, the site was eliminated from all analyses. This treatment resulted in each site being represented by an average of 76 individuals, with 99% of the 106,160,003 sites in the final analysis (∼ 68% of the total sites in the reference genome) being covered by 58–81 clones.

Estimation of population genetic parameters

As noted above, site-specific allele and genotype frequencies were determined using the ML procedures of Maruki and Lynch (2015). For each site, with n individuals sampled at the site, the ML allele frequency was estimated to be the value $i/(2n)$, where i is an integer, that maximizes the likelihood of the full set of quartets of reads while simultaneously estimating the site-specific error rate. Population-level LD was estimated using the ML method of Maruki and Lynch (2014), and individual-level disequilibrium (correlation of zygosity) using the methods of Lynch (2008) as implemented in Haubold *et al.* (2010). The methods outlined in Maruki and Lynch (2014, 2015) are easy to implement, requiring as input the simple quartets of reads (numbers of inferred A, C, G, and T nucleotides) at each site in each individual, and factor out all sources of sequencing errors (not simply quality scores); for the mean sequence coverages and numbers of individuals deployed in this study, these methods yield estimates that are essentially unbiased with sampling variance close to the minimum theoretical possibility, and hence perform as well as (and in many cases better than) other existing methods.

Data availability

All of the data utilized in this project are publicly available. The genomic assembly for *Da. pulex* PA42 v3.0 is available at the European Molecular Biology Laboratory (EMBL) nucleotide sequencing database under accession PRJEB14656. The raw reads for the KAP-population genotypes have been deposited in the NCBI Sequence Read Archive (SRA) under accession SAMN06005639, and the *Da. obtusa* genome sequences can be accessed at the NCBI SRA under accession SAMN06013355.

Results

Deviations from Hardy–Weinberg expectations

Greater than 98% of sites with minor allele frequency (MAF) estimates > 0.03, and essentially all sites with MAF > 0.04, were deemed to be significant polymorphisms at the 0.05 probability level (Figure 1 and Supplemental Material, File S1). The average inbreeding coefficient across all sites with MAF > 0.04 is 0.014, implying an average genome-wide 1.4% deficit of heterozygotes relative to Hardy–Weinberg expectations. Approximately 14% of sites with MAF > 0.15 exhibit significant deviations from Hardy–Weinberg expectations (at a probability level of 0.05); the shift in statistical behavior below this point is simply a matter of inadequate statistical power, as the expected number of observed MAF homozygotes is < 2.0 below MAF 0.15.

Rather than being a direct product of annual inbreeding (which in principle might occur via rare within-clone matings), this pattern of slight homozygote excess might be an indirect consequence of the recruitment of annual hatchlings from resting eggs deposited in multiple years, which naturally leads to a heterozygosity deficit if allele frequencies change through time (Templeton and Levin 1979). Alternative explanations that cannot yet be ruled out are the possibility that rare, segregating gene-sized deletions lead to an artifactual appearance of the small homozygote excess, and that a small amount of mapping bias might reduce the incidence of reads containing an excess of variable sites.

To put the slight deficit of heterozygotes in the KAP population into perspective, it is desirable to draw comparisons with results for other obligately sexual species. Unfortunately, the substantial amount of population genomic inquiry in *Drosophila* has consistently been performed on lines that have been intentionally inbred in the laboratory. Nonetheless, earlier work on allozymes in wild-caught flies implies that such populations typically exhibit heterozygote deficits equal to or greater than that observed in *Da. pulex*. For example, in a study of four loci in 17 populations of *D. melanogaster*, Johnson and Shaffer (1973) found an average degree of local inbreeding of 0.037 (SE = 0.017), with two of four loci exhibiting highly significant heterozygote deficit. A subsequent study of eight polymorphic loci in dozens of North Carolina populations yielded an average $F_{is} = 0.033$ (Smith *et al.* 1978). Likewise, in a study of six loci in 13 independent samples of *D. buzzatii*, Prout and Barker (1993) found an average $F_{is} = 0.096$, with a lower 95% C.I. of 0.033. Thus, despite its periodic clonal nature,

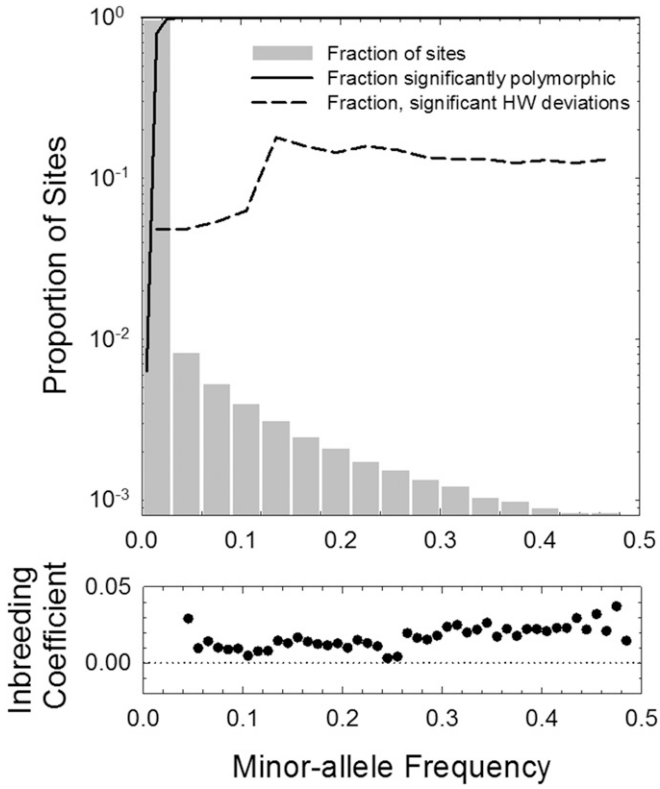

Figure 1 Top panel: Summary of the polymorphism and Hardy–Weinberg (HW) statistics, with fractions of sites with different minor allele frequency (MAF) estimates given as the bar graph. Bottom panel: The average inbreeding coefficient, F_{is}, as a function of the MAF. The dotted line is the expected value (0.0) under random mating.

the study population exhibits less inbreeding than a typical *Drosophila* population.

Nucleotide diversity

Nucleotide diversities, measured as the average expected heterozygosity over sites (π, under the assumption of Hardy–Weinberg equilibrium) and obtained using the formula of Tajima (1983), exhibit considerable variation across the genome (Table 1). The average nucleotide heterozygosity across the entire genome is 0.0094, with silent sites and sites within a restricted region of coding introns (see below) having the highest levels of variation: 0.0183 and 0.0173, respectively. Replacement sites within protein-coding sequence have by far the lowest diversity, 0.0031, whereas variation within UTR exons and the deeper intergenic regions ($>$600 bp from start and stop codons) falls in the narrow range of 0.008–0.010 (with that for UTR introns being slightly elevated).

There is considerable spatial variation of nucleotide diversity across known coding regions. For example, with respect to coding region introns, π is strongly depressed in the first five (5′-end) and last two (3′-end) nucleotide positions, likely because these sites are involved in splice-site recognition (Figure 2). Average nucleotide diversity within introns then remains at a high level at both ends up to internal

Table 1 Nucleotide diversities per site (π) and nucleotide usage (given as proportions)

Genomic region	$\hat{\pi}$	A	C	G	T
Total	0.0094	0.296	0.204	0.204	0.297
Intergenic	0.0090	0.298	0.201	0.201	0.299
Head-to-tail	0.0096	0.297	0.202	0.202	0.298
Head-to-head	0.0089	0.297	0.203	0.203	0.297
Tail-to-tail	0.0084	0.305	0.196	0.195	0.304
Coding exon	0.0076	0.274	0.246	0.235	0.245
Silent site	0.0183	0.230	0.278	0.207	0.285
Replace site	0.0031	0.298	0.215	0.265	0.223
5′-UTR exon	0.0087	0.286	0.212	0.206	0.297
3′-UTR exon	0.0078	0.293	0.200	0.182	0.324
Coding intron	0.0119	0.284	0.190	0.183	0.343
Restricted	0.0173	0.295	0.154	0.124	0.427
5′-UTR intron	0.0109	0.282	0.193	0.194	0.331
3′-UTR intron	0.0121	0.293	0.184	0.180	0.343

Silent-site estimates are based on fourfold redundant sites in codons, whereas replacement sites are zero-fold redundant sites. Full intron analyses exclude the first and last 10 bp, whereas restricted intron analyses span positions 8 to 34 from both ends; and intergenic analyses exclude the first and last 600 bp between the start and stop codons of adjacent genes.

positions $\pm$34, thereafter declining gradually to a constant level beyond $\sim$ 150 bp from both intron ends. This pattern is essentially symmetric at both ends of introns. As the modal intron size is 66 bp (Figure 2), with only a small fraction being longer than 100 bp, the vast majority of intron sites reside within the plateau of maximum diversity.

Silent-site diversity within coding exons is slightly depressed within the first five bases adjacent to introns, suggesting involvement in splice-site recognition, then remaining quite constant and similar to that seen in the high-diversity regions of introns (Figure 2); although there is a slight further increase in π_s more deeply into coding exons (beyond $\sim$ 150 bp from the intron–exon junction), only a tiny fraction of exons exceed 300 bp in length. In the first and final coding exons, there is depressed silent-site diversity in the immediate ($\sim$ 150 bp) vicinity of the translation initiation and termination codons, whereas nucleotide diversity is nearly constant across the entire intergenic region, including that incorporated into UTRs (Figure 3).

It is useful to contrast the preceding results with those for the one other well-established model arthropod system for population genomics, *D. melanogaster*. For purposes of comparison, we confine discussion to the autosomes of *D. melanogaster*, as *Daphnia* do not harbor sex chromosomes. Drawing from results in Campos *et al.* (2012), Mackay *et al.* (2012), and Huang *et al.* (2014), the average site-specific heterozygosity in *D. melanogaster*, 0.0056, is $\sim$ 60% of that in *Da. pulex*, and reductions are seen in all regions: the ratio of the former to the latter is 0.76 for fourfold redundant sites; 0.50 for zero-fold redundant sites; 0.68 for introns; 0.42 for UTR regions; and 0.69 for intergenic regions.

Triallelic and tetra-allelic sites

Although all of the preceding and subsequent analyses relied upon the subset of sites that were either monomorphic

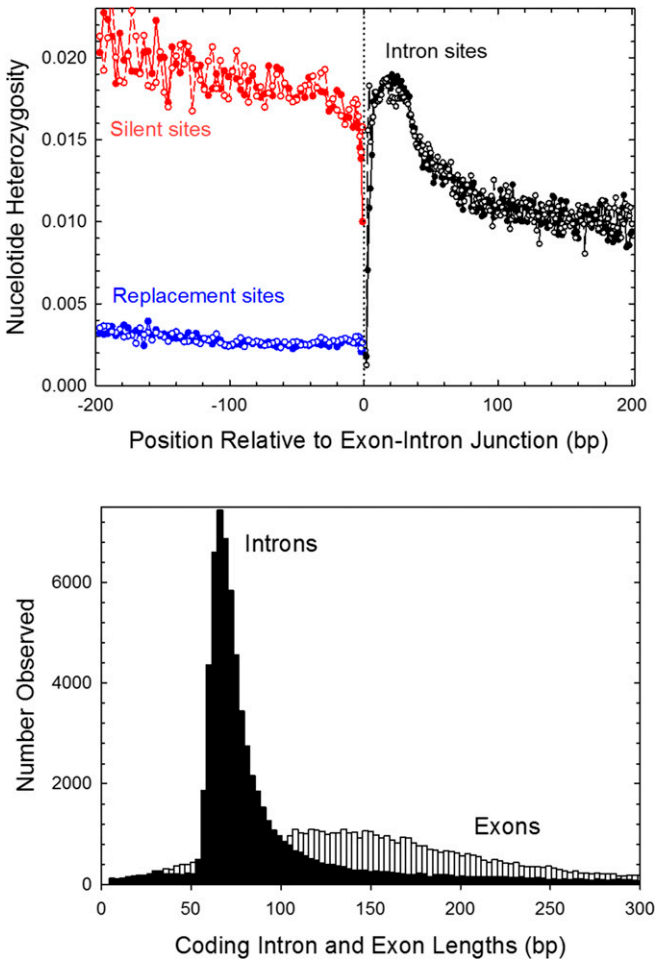

Figure 2 Top panel: Position-specific nucleotide diversities (π) averaged over all coding region introns and exons, with position 1 denoting the first (5′-end) or final (3′-end) site within an intron, *i.e.*, the measurement scale is reflected at both ends of the intron to demonstrate the similar patterns at the 5′- (closed points) and 3′- (open points) ends. Only internal introns and exons (coding, but devoid of translation initiation and termination sites) are included. In the analyses for each intron/exon, calculations were truncated at the midpoint so that data are nonoverlapping, *i.e.*, for an exon of length x, data from each end were confined to the $x/2$ adjacent sites. The vertical dotted line denotes the exon–intron junction. Bottom panel: Frequency distributions of sizes of internal coding region introns and exons.

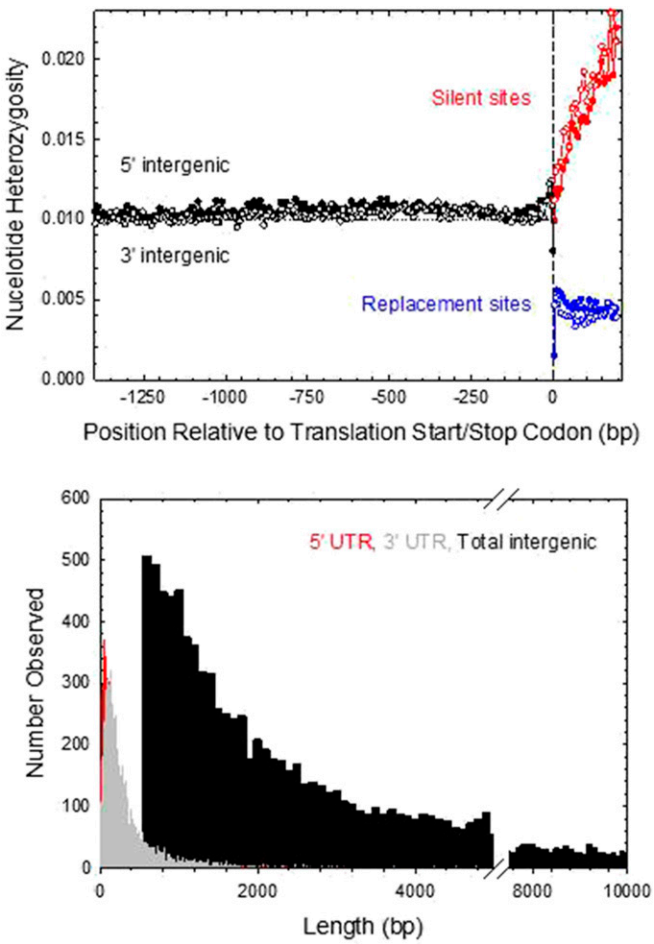

Figure 3 Top panel: Position-specific nucleotide diversities (π) averaged over all intergenic regions, here defined as the span between translation initiation and termination codons, regardless of orientation, and for the adjacent first and final coding exons, with position 1 (vertical dashed line) denoting the first position in the initiation codon or final position in the termination codon, *i.e.*, the measurement scale is reflected at both ends of the coding region to demonstrate the similar patterns at the 5′- (closed points) and 3′- (open points) ends of the gene. The horizontal dotted line simply denotes the position where $\pi = 0.01$ for reference. Bottom panel: Frequency distributions of lengths of total intergenic regions (black, and confined to regions >500 bp beyond translation, initiation, and termination codons) and the portions known to be transcribed into UTRs. Much of the 5′-UTR distribution is hidden from view, but the overall distribution is very similar to that for 3′-UTRs.

or biallelic, using a ML procedure to explicitly identify individual genotypes (T. Maruki and M. Lynch, unpublished results), a small fraction of sites was found to contain three or four variant nucleotides. Applying this procedure to 106,493,146 sites in the population sample, and using a 0.05 significance level as a cutoff for genotype calling, the fractions of sites containing two, three, and four nucleotides were 0.0482, 0.00137, and 0.0000216, respectively. Comparing these observations with the expectations from a model that allows for multiple mutations on a site within a sampled genealogy, but corrects for parallel mutations (File S1), sites segregating three nucleotides are 2.2× more frequent than expected by chance, and the incidence of tetra-allelic sites is inflated by a factor of 4.0 (Figure 4). Using a different type of analysis, Hodgkinson and Eyre-Walker (2010) found a 1.8× inflation of triallelic sites in human genome sequences relative to expectations under a null model of random mutation, but there appear to be few other recorded data on this issue.

Strength of selection on molecular polymorphisms

The similar and relatively high levels of nucleotide diversity for both fourfold redundant silent sites and sites within the restricted region of introns noted above suggest that these two subclasses may be evolving in a nearly neutral fashion. This possibility can be evaluated more closely by examining the site frequency spectrum (SFS) (Figure 5A). For a set of

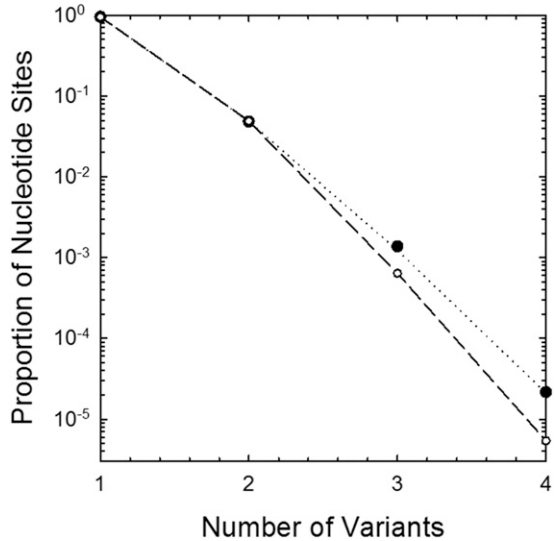

Figure 4 Observed frequencies of sites segregating 1, 2, 3, and 4 nucleotides, contrasted with the expectations from a model assuming independent mutations and correcting for parallel mutation (dashed line and open points; see File S1). The dotted line is the uncorrected Poisson distribution.

sites under purifying selection against mutations with additive deleterious effect s, the folded site frequency distribution has the form

$$p_s(x) = C \cdot \frac{2 - e^{Sx} - e^{S(1-x)}}{x(1-x)}, \tag{1a}$$

where $p_s(x)$ denotes the frequency of minor alleles in allele frequency class $x \leq 0.5$, and $S = 4N_e s$ is twice the selection coefficient scaled to the power of genetic drift $1/(2N_e)$. The coefficient $C = k(4N_e u)/(1 - e^S)$, where u is the mutation rate per nucleotide site, N_e is the effective population size, and k is a normalization constant that depends on the width of the frequency classes (necessary to ensure that the site frequencies sum to 1.0) (Wright 1938; Messer 2009). Under neutrality, Equation 1a reduces to

$$p_n(x) = k \cdot \frac{4N_e u}{x(1-x)}, \tag{1b}$$

with k again being the normalization factor. Up to MAF 0.35, the inverse scaling with $x(1-x)$ expected under neutrality is closely approximated with both fourfold redundant silent sites and sites within the restricted region of introns (Figure 5A).

It is clear from Figure 5A that the fraction of coding region, replacement site mutations that are able to increase to high frequency is much reduced, whereas the situation is intermediate for sites within UTRs and intergenic regions. All of the latter categories of sites are almost certainly heterogeneous with respect to fitness effects, with some sites likely being under strong selection and others being nearly neutral, as can be seen by the fact that most of the drop in the SFS for all three categories occurs in the lowest frequency class (average MAF = 0.015).

That there is a class of mutations in all of these categories of DNA is under substantially stronger selection than those residing in other sites, can be seen by the following argument. By comparing the abundance of a particular frequency class for selected (s) sites against that observed for neutral (n) sites, using the ratio of Equations 1a and 1b, one can acquire an estimator of the strength of selection that eliminates the scaling and mutational terms $k \cdot (4N_e u)$,

$$\frac{p_s(x)}{p_n(x)} = \frac{2 - e^{Sx} - e^{S(1-x)}}{1 - e^S}. \tag{2a}$$

For amino acid replacement sites, the frequency of minor alleles in class $x = 0.015$ is 0.00253, whereas that for neutral sites is 0.01070, which, according to Equation 2a requires $S \simeq 96$, a very strong level of selection. If all mutations experienced this level of selection, $p_s(x)/p_n(x)$ would equal 7×10^{-5} at an allele frequency of $x = 0.1$, about three orders of magnitude below what is actually seen (Figure 5A).

Thus, there must be a substantial pool of mutations with more mildly deleterious (to neutral) effects to account for the fact that, in regions under selective constraint, almost all of the drop in the SFS is incurred in the first few frequency classes. As selection weeds out the most deleterious alleles relatively rapidly, only those within the subset behaving in an effectively neutral fashion are expected to rise to high frequency. If this is the case, it follows from the above formulae that $p_s(x)$ and $p_n(x)$ should decline at the same rate with increasing x, and this is approximately the case for sites with MAF > 0.05 for replacement, UTR, and intergenic sites (Figure 5B). Thus, the majority of SNPs that have risen to frequency 0.05 are likely to be effectively neutral.

To gain more insight into the distribution of fitness effects associated with mutations in various classes of genomic sequence, we have utilized an approach applied in prior studies (Eyre-Walker and Keightley 2007; Keightley and Eyre-Walker 2010). As neither a model with a single S or a unimodal distribution of S will adequately fit the data, we assumed the presence of one class of mutations with a major and essentially constant deleterious effect S_M combined with a continuous distribution of more minor effects S_m following a gamma distribution,

$$\phi(S_m) = \frac{\beta^\alpha S_m^{\alpha-1} e^{-\beta S_m}}{\Gamma(\alpha)}, \tag{2b}$$

with mean α/β, variance α/β^2, and coefficient of variation $1/\sqrt{\alpha}$ for S_m. The gamma distribution is quite flexible in that it can take on forms ranging from L-shaped to nearly Gaussian depending on the relative values of α and β. Integration of Equation 2a over the distribution $\phi(S_m)$ leads to the series solution,

$$\frac{p_s(x)}{p_n(x)} = \beta^\alpha \cdot \sum_{n=0}^{\infty} \Big[(n+\beta+x)^{-\alpha} + (n+\beta+1-x)^{-\alpha} - 2(n+\beta+1)^{-\alpha} \Big]. \tag{2c}$$

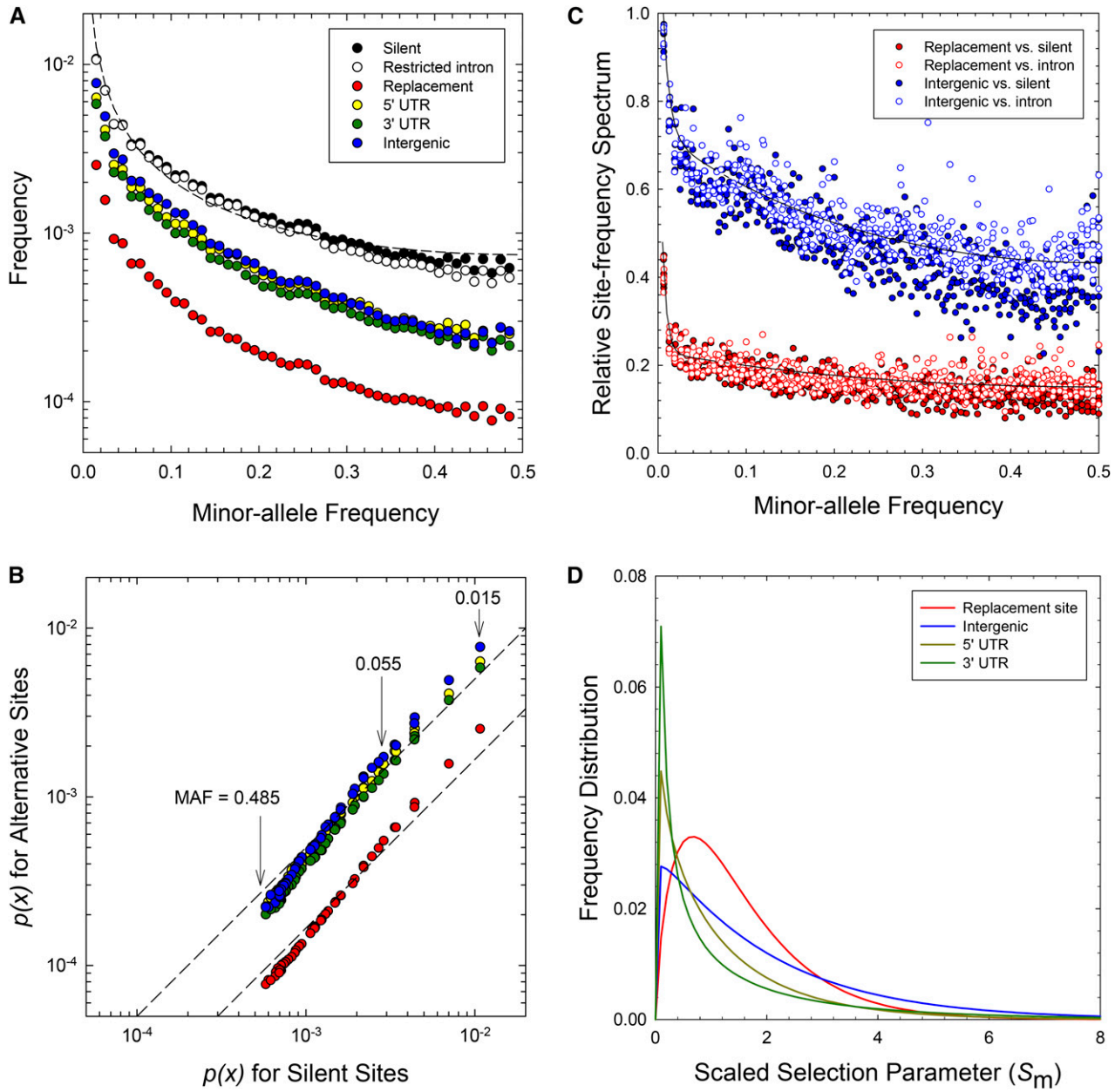

Figure 5 (A) Site frequency spectra for various classes of sites in window sizes of 0.01. The dashed line is the regression fit for the neutral expectation, Equation 1b, yielding $4kN_eu = 0.000186$. The intron analyses are confined to those not flanked by exons containing translation initiation or termination sites; UTR analyses are confined to exonic DNA. (B) Plots of the elements of the site frequency spectra for various classes of sites *vs.* those for neutral sites (taken from the previous panel). The diagonal dashed lines have slopes equal to 1.0, which is the expected behavior under neutrality. MAF denotes minor allele frequency. (C) Fitted distributions for the ratios of site frequency spectra for selected *vs.* neutral sites. Data points are from 12 comparisons, with sample sizes (numbers of individuals with adequate coverage) from 75 to 80, and silent-site and intron-site data as comparators; fitted parameters are in Table 2. (D) Fitted gamma distributions of scaled fitness effects (S_m) for the four types of selected sites, also using parameters given in Table 2.

No more than the first five terms in this expression are generally required, and often just the leading term is adequate.

Letting p_M denote the fraction of mutations in the category with major effects under this model, the overall site frequency distribution relative to the neutral expectation is

$$\frac{p_s(x)}{p_n(x)} = p_M \cdot \left(\frac{p_s(x)}{p_n(x)}\right)_M + (1 - p_M) \cdot \left(\frac{p_s(x)}{p_n(x)}\right)_m, \qquad (3)$$

where the two ratios to the right are given, respectively, by Equations 2a and 2c). Weighted least-squares fits of the observed frequency distributions relative to their respective neutral expectations (performing separate analyses for silent and intronic sites) were obtained for the SFS involving 75, 76, 77, 78, 79, and 80 individuals, and the resultant parameter estimates for p_M, S_M, and the mean and coefficient of variation of S_m were obtained by averaging the estimates for each of the 12 site frequency distributions (Figure 5C and Table 2).

Table 2 Fitted parameters for the distributions of deleterious fitness effects for genomic regions containing sites under selection

Genomic region	P_M	S_M	$\bar{S}_m$	CV(S_m)
Replacement sites	0.70 (0.03)	205 (5)	1.54 (0.14)	0.75 (0.02)
Intergenic	0.62 (0.02)	140 (4)	2.03 (0.12)	0.98 (0.03)
5′-UTR	0.48 (0.04)	204 (15)	1.24 (0.12)	1.10 (0.06)
3′-UTR	0.46 (0.04)	154 (9)	1.13 (0.12)	1.71 (0.12)

SEs are given in parentheses.

These analyses suggest that 46–70% of mutations in replacement sites in protein-coding genes, in intergenic regions, and in UTR sequences are strongly opposed by purifying selection, with S_M in the range of 140–205 (Table 2). Letting $N_e = 800,000$ (below), this implies a selective disadvantage of $s \simeq 4.4$–6.4×10^{-5} for a substantial fraction of mutations in these categories of sites. The remaining pools of deleterious mutations follow gamma distributions with far milder effects, with means in the range of $S_m \simeq 1.1$–2.0 and coefficients of variation in the range of 0.8–1.7; such distributions are highly asymmetric with long tails to the right, but with very few values exceeding $S_m = 6$ (Figure 5D). For replacement sites in protein-coding genes, drawing from data on fully sequenced *D. melanogaster* genomes, Keightley *et al.* (2016) inferred a common class of mutations ($\sim$ 90%) with $S_M \simeq 700$, with the remaining 10% having average effects of $\bar{S}_m \simeq 8$, implying approximately four- to fivefold stronger selection in this species than in *Da. pulex*.

Effective population size

Under the assumptions that silent sites and restricted intron sites are evolving in a neutral fashion, and hence only subject to the forces of mutation and drift, the standing level of heterozygosity at such sites has expected value $4N_eu$. *Da. pulex* has been subject to long-term mutation accumulation experiments, which provide a direct estimate of the base-substitution mutation rate $u = 5.69\ (1.06) \times 10^{-9}$/site/generation (Keith *et al.* 2016). Thus, taking 0.0178 (Table 1) as the average of two estimates of $4N_eu$ and scaling out $4u$, we obtain an estimate of $N_e \simeq 782,000$. As noted above, the average estimate of silent-site heterozygosity in *D. melanogaster* is 0.0139, and with an estimate of u also available for this species via a mutation accumulation study, 5.17 (0.20) $\times 10^{-9}$/site/generation (Schrider *et al.* 2013), this would suggest an effective population size of $\sim 672,000$. However, Lawrie *et al.* (2013) suggest that 22% of silent sites are under strong purifying selection in *D. melanogaster*, so if it is assumed that such sites contribute negligibly to π_s, the previous estimate needs to be increased to 862,000. In either case, it is clear that the power of both mutation and long-term random genetic drift are roughly similar in these two species.

Although the preceding estimates assume constant N_e, further insight can be acquired by considering the approximate temporal series of events necessary to account for the SFS of neutral variation, using Watterson's (1975) expression for the expected SFS for neutral sites to derive cumulative estimates of the effective population size. The rationale here is that for a given number of sampled individuals (N) and sites (L), the number of sites with i copies of the minor allele (L_i) provides a unique estimate of the population parameter $\theta = 4N_eu$. Given the availability of an estimate of u, and using the folded SFS, the estimator for each sampling interval can be evaluated as

$$\widehat{N}_{e,i} = \frac{i(2N-i)L_i}{8NLu}. \tag{4a}$$

Because the SFS strictly applies to discrete sample sizes, owing to variation in the sequence coverage at different sites, it is necessary to apply this expression separately for the pools of sites with different values of N, but this also provides a set of independent $N_{e,i}$ estimates and a simple empirical basis for obtaining SEs.

Further historical resolution can be obtained by noting that each $N_{e,i}$ estimate derived from Equation 4a reflects the full sequence of events since the origin of SNPs in that class. Taking $\hat{N}_{e,i}$ to estimate the harmonic mean of the previous interval-specific N_es (denoted N_1 for the singleton class, N_2 for the doubletons, etc.), the time sequence of interval-specific N_i estimates can then be deconvoluted from the harmonic mean estimates. Starting with the singleton class, $\hat{N}_{e,1} = \hat{N}_1$, the deconvolution formula is

$$\widehat{N}_i = \left(\frac{i}{\widehat{N}_{e,i}} - \sum_{j=1}^{i-1}\frac{1}{\widehat{N}_j}\right)^{-1}. \tag{4b}$$

Finally, the interval-specific ages (in generations) can be approximated by use of the expression of Kimura and Ohta (1973) for the ages of neutral variants,

$$t_i = -\frac{4N_{e,i}p_i \ln(p_i)}{1-p_i}, \tag{5}$$

with $p_i = i/(2N)$.

Numerous other methods have been suggested for estimating historical changes in population sizes (Strimmer and Pybus 2001; Hayes *et al.* 2003; Li and Durbin 2011; Schiffels and Durbin 2014; Liu and Fu 2015). However, aside from its considerable ease of application, this method-of-moments approach has several advantages: (1) no assumptions about the mode of population growth are required; (2) there is no need to phase haplotypes, nor any need for assumptions about the features of recombination; and (3) sequence gaps have no influence on the outcome. However, like all methods for reconstructing demographic history, the current approach becomes increasingly unreliable when extended to long time series because of the diminishing frequency of old polymorphisms, and also because small variation in harmonic-mean estimates resulting from sampling errors starts to translate into very large fluctuations in the interval-specific estimates necessary to fit the model. The lower limit to the timescale that can be analyzed is set by the expected age of singleton mutations.

We applied the preceding estimator to the SFS for five pools of observations based on $N = 75$–79 individuals, all of which contain at least 10^5 sites. Separate analyses were performed for fourfold redundant sites in coding regions and for the restricted set of neutral sites in introns noted above. For up to eight intervals (site frequency classes spanning $\sim 400,000$ generations), fairly consistent results are obtained with both silent-site and intron data, suggesting a slow decline of the effective population size from $\sim 1,200,000$ to $\sim 500,000$ at a period $\sim 60,000$ generations ago (Figure 6). It is worth bearing in mind that this time span extends prior to the founding of the study population, as the study pond is in a disturbed landscape and probably < 150 years old), so that the population size estimates are more indicative of the metapopulation level N_e.

As a check on these results, we applied to several clones the Pairwise Sequentially Markovian Coalescent (PSMC) method of Li and Durbin (2011), which draws information from the homozygosity-tract length distribution based on all sites, and the stairway method of Liu and Fu (2015). The latter method, which was applied specifically to our observed silent-site SFS, is much more computationally intensive, as it evaluates significance levels of differences in the interval-specific N_i. Our results are intermediate to those obtained by both methods, with the PSMC uniquely predicting a decline in population size in the deep past, which is perhaps a consequence of bias introduced by the inclusion of selected sites in the analysis. Despite the qualitative differences, all three methods imply an $\sim 50\%$ decline in N_e over the past $\sim 300,000$ generations.

These results permit closer scrutiny of the results derived from the SFS in the previous section, which relied on the assumption of a stable N_e. To evaluate the SFS under an unstable population size, we used the software $\partial a \partial i$ (Gutenkunst *et al.* 2009) to generate expected SFS for various values of $S = 4N_a s$, where N_a is the ancestral effective population size, here assumed to be 10^6, with N_e then undergoing an exponential decline to 500,000 starting 250,000 generations in the past (approximating the average results in Figure 6). As can be seen in Figure 7, $p_s(x)/p_n(x)$ under this demographic scenario is virtually identical to that under the assumption of constant N_e for $S \leq 2$, and only slightly elevated for S as large as 8. As the fitted distributions for S_m in Figure 5D have almost their entire density below $S = 8$, this suggests that for this particular study population, violations of the assumption of constant N_e will have minor effects on inferences on the distribution of fitness effects. For very large S, the disparity in the SFS under stable and declining population sizes becomes greater, but even with S as large as 200, the stable population size assumption will not lead to more than an $\sim 30\%$ underestimate of the true S, so the conclusions reached above regarding the magnitude of S_M are also qualitatively upheld. There are, of course, some uncertainties in these calculations owing to the variation in demographic scenarios depicted in Figure 6, but it does not appear that a twofold reduction of N_e on the time scale observed is sufficient to alter the conclusions reached above.

Nucleotide usage

As can be seen from Table 1, the *Da. pulex* genome is highly AT-rich, particularly in the restricted region of introns harboring high levels of variation. The coding regions exhibit less strong AT bias, although there is still disparity in the usage of alternative nucleotides at both silent and replacement sites. As noted above, *Da. pulex* and *D. melanogaster* have both been subject to long-term mutation accumulation experiments, and these reveal fairly similar molecular spectra of spontaneous mutations (Table 3). Given the ratio of the summed mutation rates in the direction of (C/G) → (A/T) to the reverse, 2.725 (0.217) and 2.978 (0.210), respectively in these two species, the equilibrium A/T composition of these genomes would be expected to be roughly the same, 0.732 (0.146) and 0.749 (0.143), if governed by mutation alone. No region of the *Da. pulex* genome reaches this extreme (Table 1), although A/T composition in the restricted region of coding introns (0.722) is very close to this expectation.

Given that variants at fourfold redundant sites appear to behave in an essentially neutral fashion, how can the relatively low A/T composition (0.515) of such sites be reconciled? A possible explanation relates to the fact that the spectrum of mutations is context-dependent, with the behavior of any particular site depending on the nature of the flanking bases (Long *et al.* 2014; Sung *et al.* 2015). All fourfold redundant sites are present in third positions of codons, while the flanking second positions are uniformly replacement sites and therefore generally nearly invariant, and first positions are also generally under strong purifying selection. As coding sequences tend to be GC rich, this means that the context of nucleotide sites (and hence the molecular spectrum of mutations) in such sequence differs from that in introns, where there is little selective constraint at most sites. Unfortunately, without hundreds of *de novo* mutations, which we do not have for invertebrates, it is not possible to explicitly account for such effects in deriving the expected nucleotide composition.

There are substantial differences in codon usage in *Da. pulex* and *D. melanogaster* (Table 4). For both two- and fourfold redundant sites, the most commonly used codon for all but one amino acid in *D. melanogaster* ends in C or G. In contrast, the most commonly used bases in the third positions of *Da. pulex* codons are A and T in the majority of cases, more in line with a prevailing role for mutation pressure. Given the similarities of the mutational spectra in the two species, this suggests significant differences in the strength of selection and/or biased gene conversion. In both species, C/G usage is substantially higher in silent sites than the neutral expectation, indicating that the prevailing direction (but not the strength) of selection/conversion pressure is the same.

To gain insight into the magnitudes of opposing forces necessary to account for the different codon usages in these two species, we use the expression for the expected frequency of a C/G (as opposed to A/T) nucleotide in the third position of

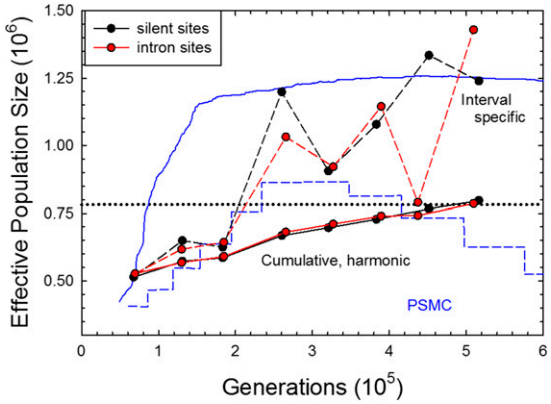

Figure 6 Historical changes in the effective population size from the present (left) to the past (right) derived from the frequency spectrum for neutral sites (black and red). Separate results are given for estimates derived from silent sites and restricted intron sites. The cumulative harmonic-mean population sizes are obtained directly from Equation 3, and the interval-specific estimates are those required to account for the cumulative harmonic means. The horizontal dotted line denotes the average estimate of N_e derived from the standing levels of heterozygosity for these sites. The lower blue plot denotes population-size estimates derived from the estimated homozygosity-tract length distribution using the method of Li and Durbin (2011), whereas the upper blue curve is derived by the method of Liu and Fu (2015). Assuming five generations per year, 200,000 generations implies 40,000 years.

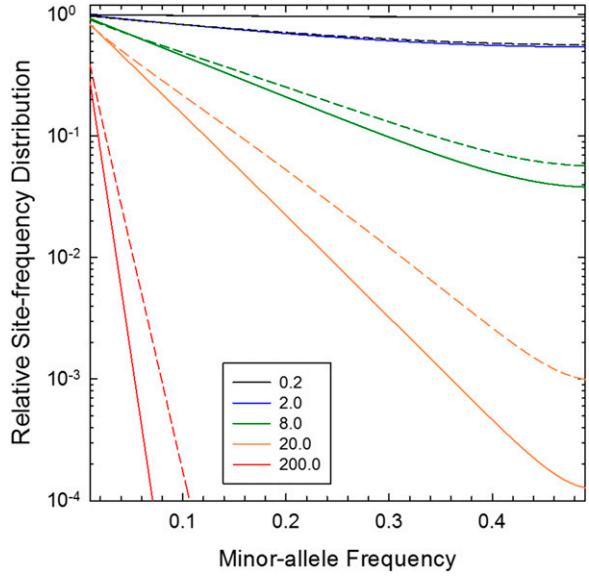

Figure 7 Comparisons of the expected site frequency spectra for deleterious mutations (scaled to the expected site frequency spectrum for neutral sites) with various values of $S = 4N_a s$ (denoted in the key) in the ancestral population, with the alternative demographic scenarios outlined in the text. Solid lines denote the results under a stable N_e, and dashed lines denote those under a scenario of a twofold decline over the past 250,000 generations.

amino acids within two-codon sets (all of which utilize transition pairs G/A or C/T),

$$\tilde{p}_{C/G} = \frac{me^S}{me^S + 1}, \tag{6}$$

where m is the ratio of A/T $\leftrightarrow$ C/G mutation rates in the direction of C/G *vs.* A/T, and $S = 4N_e s$, with N_e being the effective population size, and s the selection coefficient in favor of C/G codons (Li 1987; Bulmer 1991). This expression assumes that nucleotide usage has settled into long-term drift-mutation-selection equilibrium, and that selection is strong enough to generally maintain the amino acid identity of the sites under consideration. Application of Equation 6 shows that, in both species, selection always favors C/G-ending codons ($S > 0$), and that S is on the order of 1.0–2.0 for all amino acids in *D. melanogaster*, but consistently smaller in *Da. pulex* (Table 4). The average estimate of $S = 1.63$ (0.13) obtained here for *D. melanogaster* is similar to an earlier estimate of 1.33 obtained by different methods (Zeng and Charlesworth 2010). From the data in Table 4, for the third positions in two-codon amino acids, the ratio of the strength of selection to the power of drift, $s/(1/2N_e) = 2N_e s$, is $\sim 2.4\times$ higher in *D. melanogaster* than in *Da. pulex*.

It is less straightforward to estimate the features of selection with fourfold redundant sites (McVean and Charlesworth 2000), although it is again clear that, in every case, codon usage is much more biased in *D. melanogaster* than in *Da. pulex*. As a rough approximation to the problem, we will consider the most commonly used codon in *D. melanogaster* to be the optimal state, with the three others being equally disadvantageous. Treating the system as effectively having two allelic classes, and using the data in Table 4, Equation 6 can again be used to estimate the scaled intensity of selection (S) on the optimal codon (again, always ending in C or G in *D. melanogaster* as confirmed by the positive estimates of s). Such an analysis suggests that in *Da. pulex*, the average selection intensity in favor of the nucleotide identified as optimal in *D. melanogaster* is not significantly different from zero, consistent with the assertion made above that such sites are evolving in an effectively neutral fashion. In contrast, the mean value of S for fourfold redundant sites in *D. melanogaster*, 0.57 (0.10), is highly significant, although threefold weaker than that associated with twofold redundant sites.

Selection on coding DNA

To draw inferences about the general direction and magnitude of selection on amino acid sequences, we relied on both the gene-specific values of π_N and π_S for nonsynonymous and synonymous variation, as well as divergence measures at replacement and silent sites, D_N and D_S, based on a comparison of the two assembled *Da. pulex* genomes, one from the midwest US (PA42, used here and geographically near KAP; Z. Ye, S. Xu, K. Spitze, J. Asselman, X. Jiang, M. E. Pfrender, and M. Lynch, unpublished results) and one from the state of Oregon (TCO; Colbourne *et al.* 2011), which is known to be substantially isolated and sometimes given a different species name (*Da. arenata*) (Crease *et al.* 1997). In addition, we were able to exploit existing sequence data ($\sim 10\times$ coverage) from a clone of a somewhat further outgroup species, *Da. obtusa*.

Table 3 Conditional base-substitutional mutation rates (and their SEs) in units of $\times 10^{-9}$/site/generation, taken from Schrider *et al.* (2013) for *D. melanogaster* and Keith *et al.* (2016) for *Da. pulex*

Mutation type	*Da. pulex*	*D. melanogaster*
A:T → C:G	0.610 (0.014)	0.529 (0.082)
A:T → G:C	1.550 (0.057)	1.116 (0.119)
A:T → T:A	0.995 (0.032)	1.012 (0.114)
C:G → A:T	1.435 (0.174)	1.406 (0.159)
C:G → G:C	0.495 (0.039)	0.837 (0.122)
C:G → T:A	4.450 (0.405)	5.670 (0.318)

With the average value of π_S (based on fourfold redundant sites) for the KAP population being 0.0183 (0.0002), the TCO-PA42 D_S being 0.0488 (0.0004), and the above results suggesting neutrality at such sites, the mean coalescence time of pairs of alleles across the two *Da. pulex* populations is $\sim 2.7\times$ the average coalescence time of alleles within the KAP population. As the expected entire depth of a gene genealogy within a randomly mating population is twice the average coalescence time for randomly sampled alleles, this means that the average level of interpopulation divergence is ~35% greater than the average depth of allele genealogies within the KAP population. The average silent-site divergence between midwest *Da. pulex* and *Da. obtusa* is 0.1090 (0.0005), which implies a divergence time across species $\sim 3\times$ the average coalescence time of gene genealogies within the KAP population.

Considering only the subset of genes with D_S and π_S both >0.005 to avoid spuriously extreme ratios, which leaves 7319 genes for analysis, the average value of π_N/π_S within the KAP population is 0.158 (SE = 0.002), implying that $\sim$ 84% of amino acid-altering mutations are unable to rise to high frequencies within this population. In striking contrast, the average D_N/D_S for the TCO-PA42 comparison is 0.251 (SE = 0.003), whereas that for the *Da. obtusa*-PA42 comparison is just 0.141 (0.002). For genes within this pool, only 0.7% have $\pi_N/\pi_S > 1.0$, whereas 2.4 and 0.5% have $D_N/D_S > 1.0$, respectively, in the TCO-PA42 and *Da. obtusa*-PA42 comparisons.

Taken together, these results imply that substantially more amino acid altering mutations are able to attain high frequencies in the Oregon clade of *Da. pulex* than can achieve measurable heterozygosity within the KAP population, whereas from the perspective of the close relative *Da. obtusa* fewer replacement mutations are able to go to fixation than can segregate. Whereas one interpretation of this pattern might be that the Oregon clade has undergone a massive bout of adaptive fixation, the evidence suggests quite a different situation. The TCO genome was derived from a population known to be highly inbred (with $<10\%$ of the nucleotide diversity present in the KAP population; Colbourne *et al.* 2011), which is likely a consequence of a long-term population bottleneck in this geographic region (Lynch *et al.* 1999; Z. Ye, S. Xu, K. Spitze, J. Asselman, X. Jiang, M. E. Pfrender, and M. Lynch, unpublished results). Thus, rather than providing support for adaptive amino acid change in Oregon *Da. pulex*, the current data are much more consistent with the specific accumulation of mildly deleterious mutations in this clade. Thus, to avoid artifactual interpretations resulting from background demographic changes, in the following attempt to infer aspects of selection on protein-coding genes, all analyses utilized *Da. obtusa* as an outgroup species. Notably, an analysis of 1,544,240 silent sites in the *Da. obtusa* genome yields an average estimate of $\pi_s = 0.0180$; this genome-wide estimate, which should be equivalent to that for a regularly randomly mating population, is virtually identical to the estimate of π_s for the KAP population (Table 1).

To determine the possibility of a subset of *Da. pulex* genes being under positive selection, we evaluated a slight variant of the neutrality index (NI) of Rand and Kann (1996), defined to be the ratio of π_N/π_S to D_N/D_S. Rather than relying on direct counts of polymorphisms, this index uses the π diversity measures, as there are some uncertainties in direct counts in low-coverage data and the ML estimates of the latter are unbiased (Maruki and Lynch 2015). Values of NI <1.0 indicate excess divergence of replacement substitutions relative to the expectations based on within-population variation, and assuming the latter to be an indicator of the neutral ratio, are generally taken to imply positive selection for adaptive amino acid substitutions. In contrast, values >1.0 are generally taken to imply a predominance of purifying selection.

Using *Da. obtusa* as an outgroup, the data in this study yield NI <1.0 for 49% of the 6965 genes for which the computation could be carried out (which requires a nonzero denominator) (Figure 8). Of the 3537 genes with NI >1.0, in only 24 cases did the observed value deviate from 1.0 by more than two SEs, the maximum deviation being just 2.9 SEs, so given the multiple comparisons being made there is no compelling reason to infer excess within-population variation at replacement sites at any locus. However, for the 3428 cases with NI <1.0, NI was significantly <1.0 in 36% of the cases (9% with the observed NI deviating below 1.0 by >6 SEs, and 3% by >12 SEs). Thus, there is evidence for excess amino acid sequence divergence in a fraction of *Da. pulex/obtusa* genes.

As just noted, NI <1 is typically viewed as an indicator of positive selection for adaptive amino acid change, with $(1-\text{NI})$ being taken as a measure of the fraction of amino acid replacement substitutions that are adaptive (Charlesworth 1994; Smith and Eyre-Walker 2002). Clearly, this strict interpretation cannot be applied to *Da. pulex*, as the average value of gene-specific NI is 1.50 (0.02). Although modifications have been suggested for estimators of NI that can reduce statistical biases associated with sampling ratios of ratios (Stoletzki and Eyre-Walker 2011), none of these eliminates the issue here of average NI >1.

A possible resolution to this problem is provided by the analyses presented above, showing that there is very strong selection on a substantial subset of replacement-site mutations in *Da. pulex* (Table 2) such that π_n/π_s declines substantially in higher frequency classes as the more strongly

Table 4 Nucleotide usage in the third positions of codons for two- and fourfold redundant sites

	Da. pulex					*D. melanogaster*				
	A	C	G	T	*S*	A	C	G	T	*S*
Twofold redundant silent sites										
Asn (AA*)		0.482		**0.518**	0.93		**0.555**		0.445	1.31
Asp (GA*)		0.448		**0.552**	0.79		0.472		**0.528**	0.98
Cys (TG*)		**0.565**		0.435	1.26		**0.710**		0.290	1.99
Gln (CA*)	**0.579**		0.421		0.68	0.302		**0.698**		1.93
Glu (GA*)	**0.663**		0.337		0.32	0.331		**0.669**		1.79
His (CA*)		**0.506**		0.494	1.02		**0.600**		0.400	1.50
Lys (AA*)	**0.622**		0.378		0.50	0.301		**0.699**		1.94
Tyr (TA*)		**0.596**		0.404	1.39		**0.630**		0.370	1.62
Fourfold redundant silent sites										
Ala (GC*)	0.217	**0.342**	0.127	0.314	0.22	0.171	**0.449**	0.188	0.193	0.67
Arg (CG*)	0.221	0.290	0.185	**0.304**	−0.02	0.194	**0.415**	0.189	0.202	0.53
Gly (GG*)	**0.320**	0.296	0.119	0.266	0.02	0.287	**0.427**	0.075	0.212	0.58
Leu (CT*)	0.155	0.252	**0.314**	0.279	0.10	0.119	0.200	**0.552**	0.130	1.08
Pro (CC*)	**0.295**	0.207	0.234	0.263	−0.47	0.249	**0.332**	0.291	0.127	0.18
Ser (TC*)	0.199	0.275	0.223	**0.303**	−0.09	0.153	**0.383**	0.326	0.138	0.40
Thr (AC*)	0.246	0.248	0.217	**0.289**	−0.23	0.196	**0.379**	0.256	0.169	0.38
Val (GT*)	0.136	0.317	0.230	**0.317**	−0.33	0.108	0.235	**0.471**	0.186	0.76

Sample sizes are in the range of 138,000 to 1,640,000, so the SEs are all < 0.001. Most commonly used bases are noted in bold. A positive value for the scaled selection parameter $S = 4N_e s$ implies selection in favor of C or G.

deleterious mutations are removed by selection. For MAF classes 0.0–0.1, 0.1–0.2, 0.2–0.3, 0.3–0.4, and 0.4–0.5, the mean values of π_n/π_s are 0.220 (0.006), 0.169 (0.002), 0.155 (0.002), 0.148 (0.003), and 0.145 (0.003), respectively. This suggests a neutral approximation of the divergence ratio of ~ 0.145, as opposed to the 0.158 obtained with the full set of segregating sites. This attempt at bias correction for purifying selection on silent sites (and the associated demographic effects) is very similar to the approach advocated by Messer and Petrov (2013).

Although this neutral expectation is still higher than the average D_N/D_S ratio of 0.141 noted above, estimates of D_N/D_S for closely related species should be corrected by subtracting the average amount of divergence simply associated with segregating variation. Because such variation is associated more with silent than with replacement sites, this removal of bias elevates the average value of D_N/D_S to 0.158 (0.004). These two modifications then lead to an overall NI estimate of $0.145/0.158 \simeq 0.92$, implying that $\sim 8\%$ of amino acid altering substitutions in the *Da. pulex/obtusa* lineage have been promoted by positive selection. This is a substantially smaller fraction than the 40–90% adaptive amino acid substitutions estimated in *D. melanogaster* by various methods, but much more similar to estimates for mammals, which mostly fall in the range of 0–30% (Eyre-Walker 2006; Benger and Sella 2013; Messer and Petrov 2013).

Linkage disequilibrium

Population-level LD, measured as r^2, declines five- to sevenfold with physical distance between sites over a span of 20 kb (Figure 9). For purely algebraic reasons, the exact values of such measures depend on the range of SNP frequencies involved in the analyses, as the upper bound to r^2 is set by the allele frequencies used. Maximum values of $r^2 = 1.0$ are only possible when allele frequencies at both sites are identical (VanLiere and Rosenberg 2008), a situation that is asymptotically approached as analyses are confined to a narrower upper range of allele frequencies. Nevertheless, the qualitative pattern of decline in r^2 is similar for all categories of allele frequencies (Figure 9). Individual-based LD, the correlation of zygosity Δ, is expected to behave in an approximately parallel manner to r^2 in a randomly mating population (Lynch *et al.* 2014), and this is seen to be approximately the case, with an ~ 20-fold decline over 20 kb (Figure 9). As can be seen for the estimates of Δ from 10 individuals, there is negligible interindividual variation in this parameter when whole-genome data are applied.

The spatial pattern of LD observed in *Da. pulex* is somewhat different from that in *D. melanogaster*. For example, an analysis based on SNPs with MAF > 0.167 in *D. melanogaster* yields values of r^2 parallel to but slightly greater than those for *Da. pulex* based on SNPs with MAF > 0.20 for sites separated by fewer than 100 bp (Figure 9), but beyond that point LD decays considerably more rapidly with distance in *D. melanogaster*. Similar patterns are seen for analyses based on a fuller set of SNPs (with a wider range of MAFs) obtained from 200 inbred lines of *D. melanogaster*. Thus, the overall picture that emerges is consistently elevated LD in *D. melanogaster* for sites separated by < 100 bp, but a slower decline in LD at greater distances in *Da. pulex*.

The reasons for this difference remain unclear. Observations on Δ from clones from nine other *Da. pulex* populations are very similar to those observed here, so the KAP population cannot be viewed as a *Da. pulex* outlier (Lynch *et al.* 2014). The relationship between LD and physical distance is a function of the fraction of recombination events resulting in crossing over (x) and the average gene conversion tract length ($\bar{L}$), but whereas $x \simeq 0.1$ in both species, $\bar{L}$ is ~ 1 kb in

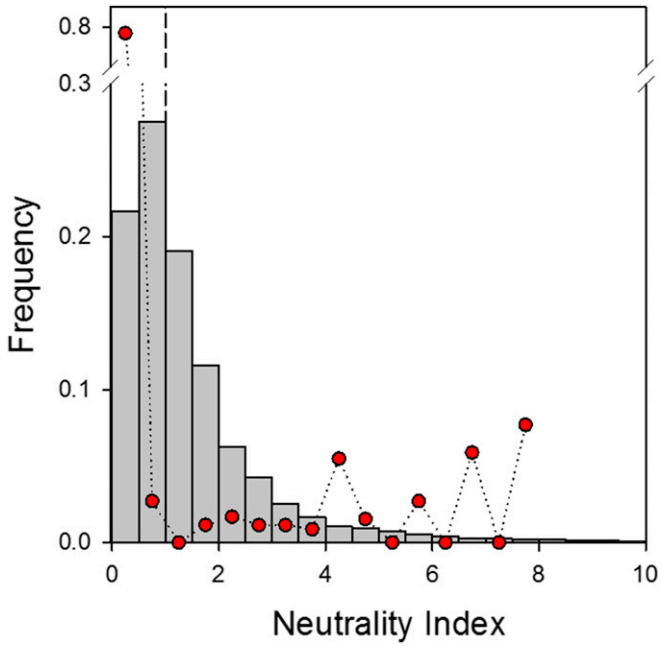

Figure 8 Distribution of the neutrality index (NI) for the 6965 genes for which the ratio is defined, with the vertical dashed line denoting the neutral expectation of 1.0, and the red dots denoting the fraction of genes within categories for which the estimate of NI deviates from 1.0 by two or more SEs.

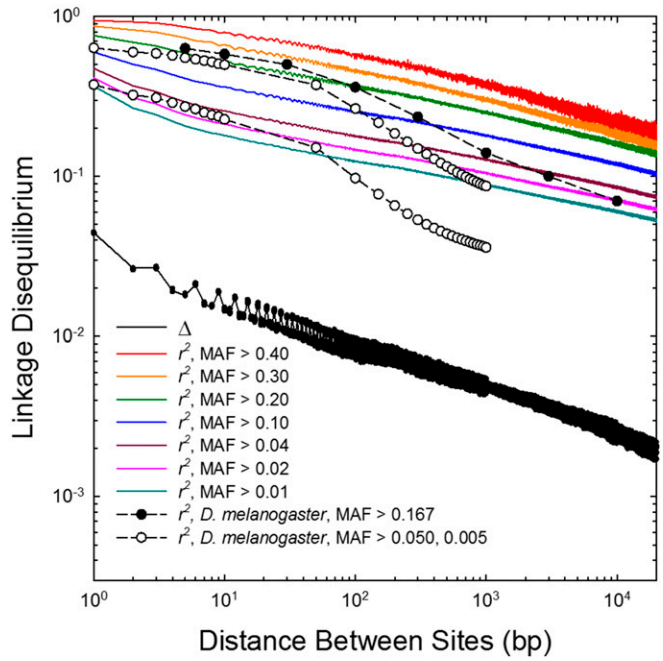

Figure 9 Measures of population-level (r^2) and individual-level (Δ) estimates of linkage disequilibrium using the full set of genomic data (partitioned into various ranges of minor allele frequencies [MAFs] in the case of r^2). For Δ, data are plotted for 10 individuals, although the differences are generally smaller than the widths of the points. The upper dashed line denotes the pattern for r^2 for SNPs with MAF >0.167 on chromosome arms 2L, 2R, and 3L, obtained from Figure 9 in Langley *et al.* (2012); the lower dashed line denotes the average pattern for the full set of SNPs obtained from both arms of chromosomes 2 and 3 from a set of 200 inbred lines (with results from two different minimal cutoffs for the MAF plotted; Huang *et al.* 2014; data provided by W. Huang).

D. melanogaster but $\sim$ 10 kb in *Da. pulex* (Lynch *et al.* 2014). The latter observation may be relevant in that shorter gene conversion tract lengths are expected to cause a more bowed pattern in the LD-distance profile, as observed in *D. melanogaster*. However, an intersection of LD-distance profiles requires that the baseline recombination rate (per generation) be lower in *D. melanogaster* (Lynch *et al.* 2014).

Although these two species have substantially different life histories and genetic systems, the average amount of recombination per physical distance does appear to be greater in temporary-pond *Da. pulex*. The latter species has 12 pairs of chromosomes, a total (male) genetic map length of $\sim$ 1450 cM, and a total genome size of $\sim$ 180 Mb (Xu *et al.* 2015), whereas *D. melanogaster* has effectively two pairs of autosomes and a pair of sex chromosomes, a female genetic map length of $\sim$ 280 cM, and a 140 Mb genome (Lindsley and Zimm 1992). *Da. pulex* is cyclically parthenogenetic, and the study population likely undergoes no more than five generations per year prior to engaging in an annual bout of sex, so the minimum number of meioses per generation is $\sim$ 0.2, whereas *D. melanogaster* lacks male recombination, which leads to 0.5 meioses per generation. Thus, assuming equal rates of recombination in male and female *Daphnia*, the amount of meiotic activity per generation in *Da. pulex* is at least 0.2×1450 cM/180 Mb $= 1.6$ cM/Mb, whereas that for *D. melanogaster* is estimated to be 1.0 cM/Mb.

The preceding results illustrate the general pattern of LD across the *Da. pulex* genome, but there is likely to be some variation in recombination rates across genomic regions. Low levels of recombination are expected to result in lower levels of heterozygosity owing to the effects of selective sweeps involving positively selected mutations at linked sites and/or background selection against linked deleterious mutations (Begun and Aquadro 1992; Charlesworth 2012). To evaluate the degree to which such expectations hold in *Da. pulex*, we scanned the entire genome for average nucleotide diversity and average r^2 between all pairs of sites within nonoverlapping windows 3 kb in length (within the range of measurable LD, as shown in Figure 9).

An interpretation of the observations requires an approximate estimate of the relative recombinational activity within genomic regions. Although it is commonly thought that the expected value of r^2 is equal to $1/(1+\rho)$, where $\rho = 4N_e c$ and c is the recombination rate between sites, Sved's (1971) derivation of this expression was based on the concept of identity-by-descent. In contrast, measures of LD with molecular markers are based on identity-in-state and have more complicated expectations that are not only functions of $\theta = 4N_e u$ (Ohta and Kimura 1971; Hill 1975) but, as noted above, have upper statistical bounds that depend on allele frequencies. More generally, the expected value of r^2 is of a form $\simeq 1/(x+\rho)$, where the value of x exceeds 1 (but is unlikely to exceed 4), varying with θ and the value of ρ itself.

The general point here is that r^2 is not a direct measure of the most desired parameter here, ρ, whereas as a transformation to $(1/r^2) - x$ is expected to roughly scale with ρ, with variation in the scaling depending on variation in the poorly known x. Thus, to gain some appreciation for the influence of recombination on standing levels of variation, we considered the relationship of the latter to $(1/r^2)$, the expectation being that π will scale positively with this surrogate measure of ρ if recombination enhances local levels of variation. Although the regression of π on $(1/r^2)$ is highly significant ($P < 10^{-6}$), and a compelling pattern is seen when the data are binned, only 5% of the variance in π is explained (Figure 10).

Discussion

Although restricted to a single population, these results establish *Da. pulex* as an excellent model for future work in ecological and evolutionary genomics. The species harbors a very high level of nucleotide diversity, and despite annual bouts of clonal reproduction, is in near Hardy–Weinberg equilibrium across the genome, a point that is consistent with considerable earlier work at the allozyme level (Lynch 1983, 1987; Lynch and Spitze 1994). Moreover, fourfold redundant sites in protein-coding genes appear to behave in an essentially neutral fashion, as do sites within regions near both ends of introns. This suggests that such sites can serve as unbiased benchmarks in studies of selection, in contrast to the situation in *D. melanogaster*, where there appears to be substantial selection on such sites (Lawrie *et al.* 2013). These two subclasses of sites will also serve as highly informative markers for evaluating aspects of effective population size and structure.

To put things in a broader context throughout, we have attempted to draw comparisons of our results with those for *D. melanogaster*, the one other invertebrate that has been subjected to a diverse array of population genetic analyses. Although there are a number of striking similarities between the population genetic features of the two species, there are also some differences, most notably an apparent elevated level of the efficiency of natural selection at the molecular level in *D. melanogaster*. For every category of genomic sites evaluated, such measures differ by factors of two to five. From the standpoint of developing generalities in empirical population genetics, it is of interest to know the degree to which such differences have a real biological basis as opposed to being artifacts of the methods of analysis that have been applied. In principle, there may be underlying ecological factors driving some of the differences; one species is aquatic, undergoes direct development, and has resting stages capable of multiple years of diapause, whereas the other has terrestrial and aerial life stages, undergoes metamorphosis, and is unknown to survive long periods in any developmental stage. However, as it is difficult to see how such differences would translate into features such as selection on silent sites and intronic sites, we seek explanations in terms of population genetic features.

The population genetic environment of *Da. pulex* is similar to that of *D. melanogaster* in several ways. First, both the mutation rate per nucleotide site per generation and the molecular spectra of mutations are very similar in both species, deviating by no more than 10%, and these features are quite compatible with those in other species with comparable effective population sizes (Lynch *et al.* 2016). Second, at least within the time span associated with standing levels of polymorphism, the long-term average effective population sizes appear to be quite similar, with that of *D. melanogaster* appearing to be $\sim$ 10% greater than that in *Da. pulex*. This conclusion is reached by treating standing levels of heterozygosity at silent sites as having an expectation equal to $4N_e u$, factoring out the mutation rate, and accounting for apparent variation-reducing selection in *D. melanogaster*.

Nonetheless, these similarities in long-term average N_e might obscure some significant underlying differences. For example, polymorphism-based methods are incapable of estimating N_e in the deep past, as even in a neutral genealogy the average time to coalescence is just $4N_e$ generations. Therefore, the possibility of larger differences in N_e of the two species at time periods greater than two million generations or so ago cannot be addressed. In addition, polymorphism-based methods for estimating the temporal history of N_e are generally unable to reveal extremely recent changes (as informative polymorphisms must have arisen to frequencies high enough to be detected in samples) nor moderately distant changes (as high-frequency polymorphisms within SFS are generally too rare to yield reliable results). Analyzed in three different ways, our results suggest an approximately twofold reduction in N_e in *Da. pulex* over the past $\sim 250,000$ generations, but our analyses suggest that this is unlikely to have greatly influenced our conclusions on selection intensity. Unfortunately, a detailed demographic analysis has not been performed on *D. melanogaster*. Although it has been suggested that this species has experienced a recent three- to fivefold (Campos *et al.* 2013; Sheehan and Song 2016) or perhaps much larger (Karasov *et al.* 2010) increase in N_e, the quantitative effects of such change on analyses of selection at the molecular level remain unclear.

The two species exhibit somewhat contrasting patterns of LD: elevated in *D. melanogaster* for pairs of sites separated by fewer than 100 bp, but decaying several-fold more slowly with further increases in physical distance in *Da. pulex*. As discussed above, these differences are compatible with a somewhat higher ($\sim$ 60%) baseline recombination rate combined with substantially longer ($\sim 10\times$) gene conversion tracts in *Da. pulex*. Nonetheless, because the pattern of decline of LD with physical distance can also be influenced by past changes in N_e (Hayes *et al.* 2003; Rogers 2014), these altered LD profiles may also partly reflect differences in demographic histories. A more sharply concave LD profile is expected for species experiencing a recent population increase than for one experiencing a population decline (Lynch *et al.* 2014), so an influence from the expected

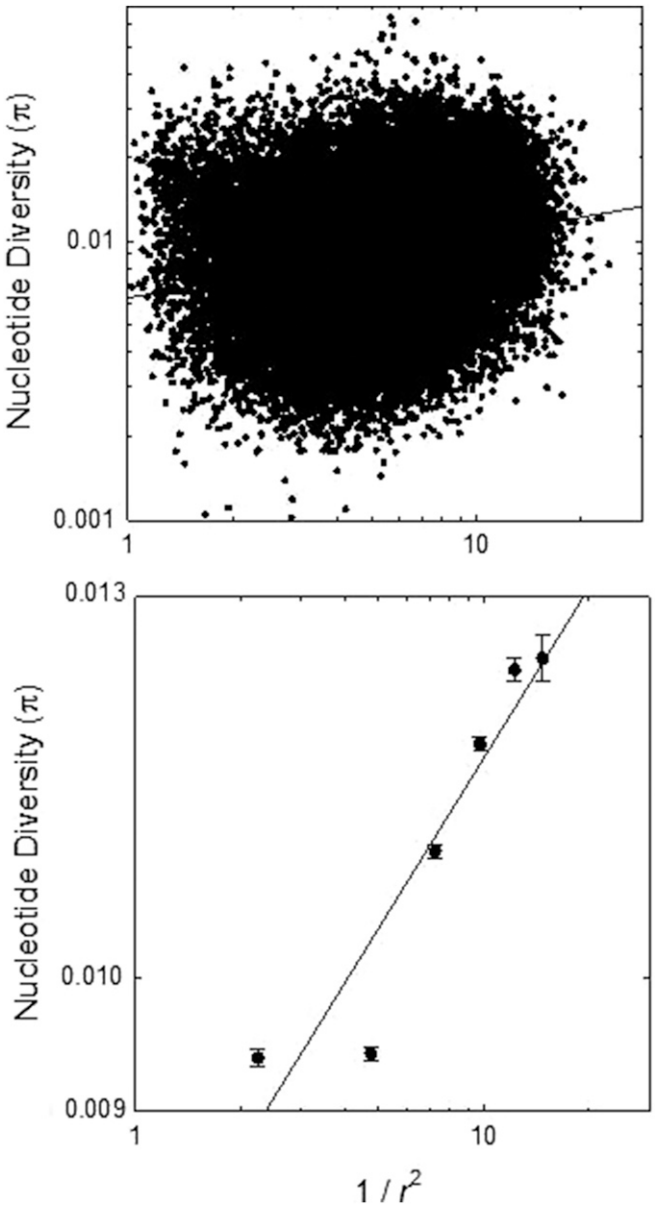

Figure 10 Relationship between average heterozygosity per nucleotide site (π) and a transformed measure of the average scaled measure of linkage disequilibrium, with each data point in the upper plot being based on results for nonoverlapping 3 kb windows across the genome. For clarity, the lower panel is based on the averages of measures within bins in the upper panel.

contrasting recent demography of *D. melanogaster vs. Da. pulex* cannot be ruled out.

As noted in Table 4, estimates of $4N_es$ for selection operating on third positions in twofold redundant codons are $\sim 2.4\times$ greater in *D. melanogaster* than those in *Da. pulex*, with the absolute disparity being potentially greater at fourfold redundant sites. Because $<10\%$ of sites involved in these analyses are polymorphic, this discrepancy cannot be viewed as an artifact of recent demographic changes in the two species. Nor can it be attributed to different methods of analysis, as data from both species were analyzed in the same way, and our results for *D. melanogaster* are compatible with prior results obtained by different methods. Thus, genome-wide differences in codon usage must be a consequence of long-term differences in the components of $4N_es$. At a minimum, either N_e in the ancient past (well beyond the mean coalescence time for current-day genes) or the absolute strength of selection s must be as much as twofold greater in *D. melanogaster*. There is no obvious way to resolve these alternatives with descriptive population genetic data, as they must have ecological underpinnings.

We can also indirectly infer a stronger efficiency of purifying selection against amino acid-altering mutations in the recent history of *D. melanogaster* than in *Da. pulex* in the following way. As noted above, silent-site heterozygosity in *D. melanogaster* is $\sim 76\%$ of that in *Da. pulex*, and if we assume that the former is depressed $\sim 20\%$ by selection on such sites (Lawrie *et al.* 2013), then the estimates of $4N_eu$ unbiased by selection are very similar. However, the genome-wide average replacement-site heterozygosity in *D. melanogaster* is $\sim 50\%$ of that in *Da. pulex*. Noting that the mean estimate of π_n/π_s in the studied *Da. pulex* population is 0.179, from Kimura's (1969) expression for the average nucleotide diversity under drift-mutation-selection equilibrium, it can be shown that a 50% reduction in π_n/π_s requires an approximately twofold increase in the strength of purifying selection ($S = 4N_es$). The more thorough analysis of the distribution of S based on the SFS provided above, which rules out a strong influence of recent demographic history in *Da. pulex* and accounts for at least some demographic history in *D. melanogaster* (Keightley *et al.* 2016), suggests that the inflation in the population-level strength of selection against deleterious amino acid altering mutations in *D. melanogaster* may be as much as $5\times$ greater than that in *Da. pulex*.

These conclusions appear to extend to other types of genomic sites and mutations. For example, from data of Andolfatto (2005), relative to the standing level of variation for silent sites, there is a substantial reduction in within-population diversity in UTRs and intergenic DNA in *D. melanogaster*, the ratios being 0.19 and 0.39, respectively, compared to values of 0.45 and 0.49 in *Da. pulex* (Table 1). In addition, as noted above, indirect inference suggests an at least fivefold increase in the rate of fixation of adaptive mutations in *D. melanogaster*. Thus, regardless of the types of sites being compared—silent and replacement sites within codons, intronic sites, and regulatory regions—it appears that purifying selection is depressing standing levels of variation whereas positive selection is promoting beneficial mutations to a greater extent in *D. melanogaster* than in *Da. pulex*.

In contrast to the analyses based on codon-usage bias, some of these polymorphism-based results may be somewhat biased by unresolved differences in the recent demographic histories of the two species. Although there is no evidence that this is an issue in *Da. pulex*, given the demographic uncertainties in *D. melanogaster*, the situation is less clear in this species. It is plausible that the absolute magnitudes of selection (the

values of s alone) are quite similar between the two species. However, for this to be the case, N_e in the deep past in *D. melanogaster* would likely have to be at least 2 × that in *Da. pulex* (to account for the differences in the fractions of fixed adaptive mutations) and any more recent population expansion in *D. melanogaster* would have to be ongoing for a sufficiently long time to leave a signature of 2 × stronger efficiency of selection on segregating sites in this species.

Should these conditions be met, then like the estimated strengths of mutation and recombination, the selection coefficients associated with molecular variants may be quite similar in these two model species. Although these kinds of uncertainties may be deemed unsatisfying, they highlight the future challenges for the field of population genomics. If generalities are to emerge, this kind of work will need to expand well beyond these two model species. Newer analytical methods allowing for varying population sizes will need to be developed, and an understanding of the mechanistic connections between ecology and genome evolution will need to be established.

Acknowledgments

We thank David Begun, Bernhard Haubold, and two reviewers for helpful comments, and Wen Huang for providing data on *D. melanogaster*. This work was supported by National Institutes of Health grant R01-GM101672 and National Science Foundation grant DEB-1257806.

Note added in proof: See Ackerman *et al*. 2017 (pp. 105–118) in this issue and Maruki and Lynch 2017 (pp. 1393–1404) and Ye *et al*. 2017 (pp. 1405–1416) in the G3 May issue for related work.

Literature Cited

Ackerman, M. S., P. Johri, K. Spitze, S. Xu, T. Doak *et al.* 2016 Estimating seven coefficients of pairwise relatedness using population genemoc data. Genetics (in press): @@@.

Andolfatto, P., 2005 Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437: 1149–1152.

Ayala, F. J. (Editor), 1976 *Molecular Evolution*, Sinauer Associates, Inc., Sunderland, MA.

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.

Benger, E., and G. Sella, 2013 Modeling the effect of changing selective pressures on polymorphism and divergence. Theor. Popul. Biol. 85: 73–85.

Bulmer, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics 129: 897–907.

Campos, J. L., B. Charlesworth, and P. R. Haddrill, 2012 Molecular evolution in nonrecombining regions of the *Drosophila melanogaster* genome. Genome Biol. Evol. 4: 278–288.

Campos, J. L., K. Zeng, D. J. Parker, B. Charlesworth, and P. R. Haddrill, 2013 Codon usage bias and effective population sizes on the X chromosome *vs.* the autosomes in *Drosophila melanogaster*. Mol. Biol. Evol. 30: 811–823.

Charlesworth, B., 1994 The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet. Res. 63: 213–227.

Charlesworth, B., 2012 The effects of deleterious mutations on evolution at linked sites. Genetics 190: 5–22.

Colbourne, J., M. E. Pfrender, D. Gilbert, W. K. Thomas, A. Tucker *et al.*, 2011 The ecoresponsive genome of *Daphnia pulex*. Science 331: 555–561.

Crease, T., S. K. Sung, S. L. Sung, N. Lehman, K. Spitze *et al.*, 1997 Allozyme and mitochondrial DNA variation in populations of the *Daphnia pulex* complex from both sides of the Rocky Mountains. Heredity 79: 242–251.

Doyle, J. J., and J. L. Doyle, 1987 A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19: 11–15.

Eads, B., D. Tsuchiya, M. Lynch, J. Andrews, and M. E. Zolan, 2012 Evolution of REC8 in *Daphnia:* the spread of a transposon insertion associated with obligate asexuality. Proc. Natl. Acad. Sci. USA 109: 858–863.

Eyre-Walker, A., 2006 The genomic rate of adaptive evolution. Trends Ecol. Evol. 21: 569–575.

Eyre-Walker, A., and P. D. Keightley, 2007 The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8: 610–618.

Frisch, D., P. K. Morton, P. R. Chowdhury, B. W. Culver, J. K. Colbourne *et al.*, 2014 A millennial-scale chronicle of evolutionary responses to cultural eutrophication in *Daphnia*. Ecol. Lett. 17: 360–368.

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

Harris, K. D., N. J. Bartlett, and V. K. Lloyd, 2012 *Daphnia* as an emerging epigenetic model organism. Genet. Res. Int. 2012: 147892.

Haubold, B., P. Pfaffelhuber, and M. Lynch, 2010 mlRho – a program for estimating the population mutation and recombination rates from shotgun-sequenced genomes. Mol. Ecol. 19(Suppl. 1): 277–284.

Hayes, B. J., P. M. Visscher, H. C. McPartlan, and M. E. Goddard, 2003 Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 13: 635–643.

Hill, W. G., 1975 Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theor. Popul. Biol. 8: 117–126.

Hiruta, C., and S. Tochinai, 2012 Spindle assembly and spatial distribution of γ-tubulin during abortive meiosis and cleavage division in the parthenogenetic water flea *Daphnia pulex*. Zoolog. Sci. 29: 733–737.

Hiruta, C., C. Nishida, and S. Tochinai, 2010 Abortive meiosis in the oogenesis of parthenogenetic *Daphnia pulex*. Chromosome Res. 18: 833–840.

Hodgkinson, A., and A. Eyre-Walker, 2010 Human triallelic sites: evidence for a new mutational mechanism? Genetics 184: 233–241.

Huang, W., A. Massouras, Y. Inoue, J. Peiffer, M. Rámia *et al.*, 2014 Natural variation in genome architecture among 205 *Drosophila melanogaster* genetic reference panel lines. Genome Res. 24: 1193–1208.

Johnson, F. M., and H. E. Schaffer, 1973 Isozyme variability in species of the genus *Drosophila*. VII. Genotype-environment relationships in populations of *D. melanogaster* from the eastern United States. Biochem. Genet. 10: 149–163.

Karasov, T., P. W. Messer, and D. A. Petrov, 2010 Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. PLoS Genet. 6: e1000924.

Kato, Y., Y. Shiga, K. Kobayashi, S. Tokishita, H. Yamagata *et al.*, 2011a Development of an RNA interference method in the cladoceran crustacean *Daphnia magna*. Dev. Genes Evol. 220: 337–345.

Kato, Y., K. Kobayashi, H. Watanabe, and T. Iguchi, 2011b Environmental sex determination in the branchiopod crustacean *Daphnia magna:* deep conservation of a Doublesex gene in the sex-determining pathway. PLoS Genet. 7: e1001345.

Kato, Y., T. Matsuura, and H. Watanabe, 2012 Genomic integration and germline transmission of plasmid injected into crustacean *Daphnia magna* eggs. PLoS One 7: e45318.

Keightley, P. D., and A. Eyre-Walker, 2010 What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? Philos. Trans. R. Soc. Lond. B Biol. Sci. 365: 1187–1193.

Keightley, P. D., J. L. Campos, T. R. Booker, and B. Charlesworth, 2016 Inferring the frequency spectrum of derived variants to quantify adaptive molecular evolution in protein-coding genes of *Drosophila melanogaster*. Genetics 203: 975–984.

Keith, N., A. E. Tucker, C. E. Jackson, W. Sung, J. I. Lucas Lledó *et al.*, 2016 High mutational rates of large-scale duplication and deletion in *Daphnia pulex*. Genome Res. 26: 60–69.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61: 893–903.

Kimura, M., and T. Ohta, 1973 The age of a neutral mutation persisting in a finite population. Genetics 75: 199–212.

Langley, C. H., K. Stevens, C. Cardeno, Y. C. G. Lee, D. R. Schrider *et al.*, 2012 Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 192: 533–598.

Lawrie, D. S., P. W. Messer, R. Hershberg, and D. A. Petrov, 2013 Strong purifying selection at synonymous sites in *D. melanogaster*. PLoS Genet. 9: e1003527.

LeBlanc, G. A., and E. K. Medlock, 2015 Males on demand: the environmental-neuro-endocrine control of male sex determination in daphnids. FEBS J. 282: 4080–4093.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*1000 Genome Project Data Processing Subgroup., 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

Li, W., A. E. Tucker, W. Sung, W. K. Thomas, and M. Lynch, 2009 Extensive, recent intron gains in *Daphnia* populations. Science 326: 1260–1262.

Li, W.-H., 1987 Models of nearly neutral mutations with particular implications for non-random usage of synonymous codons. J. Mol. Evol. 24: 337–345.

Lindsley, D. L., and G. G. Zimm, 1992 *The Genome of Drosophila melanogaster*, Academic Press, San Diego, CA.

Liu, X., and Y.-X. Fu, 2015 Exploring population size changes using SNP frequency spectra. Nat. Genet. 47: 555–559.

Long, H., W. Sung, S. F. Miller, M. S. Ackerman, T. G. Doak *et al.*, 2014 Mutation rate, spectrum, topology, and context-dependency in the DNA mismatch repair-deficient *Pseudomonas fluorescens* ATCC948. Genome Biol. Evol. 7: 262–271.

Lynch, M., 1983 Ecological genetics of *Daphnia pulex*. Evolution 37: 358–374.

Lynch, M., 1987 The consequences of fluctuating selection for isozyme polymorphisms in *Daphnia*. Genetics 115: 657–669.

Lynch, M., 2008 Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. Mol. Biol. Evol. 25: 2421–2431.

Lynch, M., and K. Spitze, 1994 Evolutionary genetics of *Daphnia*, pp. 109–128 in *Ecological Genetics*, edited by L. Real. Princeton Univ. Press, Princeton, NJ.

Lynch, M., M. Pfrender, K. Spitze, N. Lehman, D. Allen *et al.*, 1999 The quantitative and molecular genetic architecture of subdivided species. Evolution 53: 100–110.

Lynch, M., S. Xu, T. Maruki, P. Pfaffelhuber, and B. Haubold, 2014 Genome-wide linkage-disequilibrium profiles from single individuals. Genetics 198: 269–281.

Lynch, M., M. Ackerman, J.-F. Gout, H. Long, W. Sung *et al.*, 2016 Genetic drift, selection, and evolution of the mutation rate. Nat. Rev. Genet. 17(11): 704–714.

Mackay, T. F., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles *et al.*, 2012 The *Drosophila melanogaster* genetic reference panel. Nature 482: 173–178.

Mahato, S., S. Morita, A. E. Tucker, X. Liang, M. Jackowska *et al.*, 2014 Common transcriptional mechanisms for visual photoreceptor cell differentiation among Pancrustaceans. PLoS Genet. 10: e1004484.

Maruki, T., and M. Lynch, 2014 Genome-wide estimation of linkage disequilibrium from population-level high-throughput sequencing data. Genetics 197: 1303–1313.

Maruki, T., and M. Lynch, 2015 Genotype-frequency estimation from high-throughput sequencing data. Genetics 201: 473–486.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–1303.

McVean, G. A., and B. Charlesworth, 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics 155: 929–944.

Messer, P. W., 2009 Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. Genetics 182: 1219–1232.

Messer, P. W., and D. A. Petrov, 2013 Frequent adaptation and the McDonald–Kreitman test. Proc. Natl. Acad. Sci. USA 110: 8615–8620.

Möst, M., S. Oexle, S. Marková, D. Aidukaite, L. Baumgartner *et al.*, 2015 Population genetic dynamics of an invasion reconstructed from the sediment egg bank. Mol. Ecol. 24: 4074–4093.

Naitou, A., Y. Kato, T. Nakanishi, T. Matsuura, and H. Watanabe, 2015 Heterodimeric TALENs induce targeted heritable mutations in the crustacean *Daphnia magna*. Biol. Open 4: 364–369.

Nakanishi, T., Y. Kato, T. Matsuura, and H. Watanabe, 2014 CRISPR/Cas-mediated targeted mutagenesis in *Daphnia magna*. PLoS One 9: e98363.

Nei, M., and R. K. Koehn (Editors), 1983 *Evolution of Genes and Proteins*, Sinauer Associates, Inc., Sunderland, MA.

Ohta, T., and M. Kimura, 1971 Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics 68: 571–580.

Omilian, A. R., M. E. A. Cristescu, J. L. Dudycha, and M. Lynch, 2006 Ameiotic recombination in asexual lineages. Proc. Natl. Acad. Sci. USA 103: 18638–18643.

Pfrender, M. E., K. Spitze, J. Hicks, K. Morgan, L. Latta *et al.*, 2000 Lack of concordance between genetic diversity estimates at the molecular and quantitative-trait levels. Conserv. Genet. 1: 263–269.

Prout, T., and J. S. Barker, 1993 *F* statistics in *Drosophila buzzatii:* selection, population size and inbreeding. Genetics 134: 369–375.

Rand, D. M., and L. M. Kann, 1996 Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. Mol. Biol. Evol. 13: 735–748.

Rogers, A. R., 2014 How population growth affects linkage disequilibrium. Genetics 197: 1329–1341.

Schiffels, S., and R. Durbin, 2014 Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46: 919–925.

Schrider, D. R., D. Houle, M. Lynch, and M. W. Hahn, 2013 Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. Genetics 194: 937–954.

Schumpert, C., I. Handy, J. L. Dudycha, and R. C. Patel, 2014 Relationship between heat shock protein 70 expression and life span in *Daphnia*. Mech. Ageing Dev. 139: 1–10.

Schumpert, C. A., J. L. Dudycha, and R. C. Patel, 2015a Development of an efficient RNA interference method by feeding for the microcrustacean *Daphnia*. BMC Biotechnol. 15: 91.

Schumpert, C., J. Nelson, E. Kim, J. L. Dudycha, and R. C. Patel, 2015b Telomerase activity and telomere length in *Daphnia*. PLoS One 10: e0127196.

Sheehan, S., and Y. S. Song, 2016 Deep learning for population genetic inference. PLoS Comput. Biol. 12: e1004845.

Shiga, Y., R. Yasumoto, H. Yamagata, and S. Hayashi, 2002 Evolving role of Antennapedia protein in arthropod limb patterning. Development 129: 3555–3561.

Smith, D. B., C. H. Langley, and F. M. Johnson, 1978 Variance component analysis of allozyme frequency data from eastern populations of *Drosophila melanogaster*. Genetics 88: 121–137.

Smith, N. G., and A. Eyre-Walker, 2002 Adaptive protein evolution in *Drosophila*. Nature 415: 1022–1024.

Stoletzki, N., and A. Eyre-Walker, 2011 Estimation of the neutrality index. Mol. Biol. Evol. 28: 63–70.

Strimmer, K., and O. G. Pybus, 2001 Exploring the demographic history of DNA sequences using the generalized skyline plot. Mol. Biol. Evol. 18: 2298–2305.

Sung, W., M. S. Ackerman, J. F. Gout, S. F. Miller, E. Williams *et al.*, 2015 Asymmetric context-dependent mutation patterns revealed through mutation-accumulation experiments. Mol. Biol. Evol. 232: 1672–1683.

Sved, J. A., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Popul. Biol. 2: 125–141.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460.

Templeton, A. R., and D. A. Levin, 1979 Evolutionary consequences of seed pools. Am. Nat. 114: 232–249.

Toyota, K., H. Miyakawa, C. Hiruta, K. Furuta, Y. Ogino *et al.*, 2015 Methyl farnesoate synthesis is necessary for the environmental sex determination in the water flea *Daphnia pulex*. J. Insect Physiol. 80: 22–30.

VanLiere, J. M., and N. A. Rosenberg, 2008 Mathematical properties of the r^2 measure of linkage disequilibrium. Theor. Popul. Biol. 74: 130–137.

Watterson, G. A., 1975 On the number of segregation sites. Theor. Popul. Biol. 7: 256–276.

Wright, S., 1938 The distribution of gene frequencies under irreversible mutation. Proc. Natl. Acad. Sci. USA 24: 253–259.

Xu, S., M. S. Ackerman, H. Long, L. Bright, K. Spitze *et al.*, 2015 A male-specific genetic map of the microcrustacean *Daphnia pulex* based on single sperm whole-genome sequencing. Genetics 201: 31–38.

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.

Communicating editor: D. J. Begun