æ

SUPPLEMENTAL MATERIAL

The expected frequencies of sites segregating two, three, and four nucleotides. Consider an arbitrary gene genealogy in a sample, and let λ be the expected number of mutations at a site summed over the full length of branches on the genealogical tree. Here, we assume a situation in which the genealogy is shallow enough and the sample size small enough that no more that three mutations arise on the tree. This assumption should be closely approximated if the population-level mutation rate, $4N_e u \ll 1$, where N_e is the effective population size and u is the mutation rate per site per generation, and is further justified by the observation in this study that $\sim 95\%$ of sites are monomorphic.

We assume that the total number of mutational events on the tree is Poisson distributed, but the probability of observable numbers of mutations differs from that expected on the basis of mutational events because of the possibility of parallel mutations. Letting ϕ be the probability of that two mutations from the same parental allele type independently generate the same daughter nucleotide, and assuming that all mutations in the sample are derived from the same ancestral allele, the probabilities of a sample exhibiting 1, 2, 3, or 4 nucleotides at a site are

$$P(1) = e^{-\lambda},$$

$$P(2) = (\lambda e^{-\lambda})\{1 + [(\lambda\phi)/2] + [(\lambda\phi)^2/6]\}$$

$$P(3) = (\lambda^2 e^{-\lambda}/2)\{1 - \phi + [2\lambda\phi(1-\phi)/3]\}$$

$$P(4) = (\lambda^3 e^{-\lambda}/6)\{(1-\phi)^2\}$$

The terms to the left of the brackets are the uncorrected Poisson probabilities. Using the mutational-spectrum data in Table 2, $\phi = 0.49$ for *D. pulex,* and λ can be estimated from the observed frequency of monomorphic sites, allowing the expectations for the three remaining probabilities to be estimated by substitution.

Estimation of average within-population silent- and replacement-site diversities. To minimize all sources of uncertainty with respect to site status, we performed analyses on the specific sets of sites that could be unambiguously assigned as silent (synonymous, S) or replacement (nonsynonymous, N) sites. The silent-site analyses were restricted to the third positions in codons for alanine, glycine, proline, threonine, and valine, as in all five cases such sites are four-fold redundant; if the orthologous codons in the major and minor alleles were valine/threonine or glycine/threonine (an extremely rare situation), such sites were discarded because there are routes between such codons that might involve a context in which a nonredundant change has occured in the third position. All substitutions at second-site positions within codons cause amino-acid replacement changes; and we also counted all first positions in codons as replacement sites, provided neither codon being compared encoded for arginine, isoleucine, leucine, or serine, as excursions through such codons can again sometimes be ambiguous with respect to silent/replacement changes.

The maximum-likelihood methods deployed in this study yield estimates of the major- and minor-allele frequencies at each site, p and $(1-p)$ respectively, which provide an estimate of diversity for site i in a gene of

$$\widehat{\pi}_i = 2p_i(1-p_i). \tag{S1}$$

With L_S and L_N silent (S) and replacement (N) sites involved in the analysis for a particular gene, the gene-specific estimates of silent- and replacement-site diversities are

$$\widehat{\pi}_S = \sum_{i=1}^{L_S} \widehat{\pi}_{Si}/L_S, \qquad \pi_N = \sum_{i=1}^{L_N} \widehat{\pi}_{Ni}/L_N \tag{S2a,b}$$

From Nei and Roychoudhury (1974), the sampling variance of heterozygosity at the ith silent site is

$$\sigma^2(\widehat{\pi}_{Si}) = \frac{2(n-1)^2\widehat{\pi}_{Si}}{n^3}, \tag{S3a}$$

where n is the number of chromosomes sampled. Because some individuals may occasionally have low coverage, the number of sampled chromosomes can be somewhat less than twice the number of individuals sampled (e.g., for an individual with a coverage of just 1×, a very rare event in this study, only one chromosome can have been sampled). Thus, as an approximator of n, we relied on a measure of the effective number of chromosomes,

$$n_e = 2\sum_{i=1}^{n_{tot}}(1-0.5^{c_i}), \tag{S3b}$$

where c_i is the depth of sequence coverage for individual i, and n_{tot} is the total number of sampled individuals (after Maruki and Lynch 2015); the term $(2 \cdot 0.5^{c_i})$ is simply the probability that just a single chromosome has been sequenced from individual i.

Using the expression for the variance of a sum, the sampling variance of $\widehat{\pi}_S$ is

$$\sigma^2(\widehat{\pi}_S) = \left(\frac{1}{L_S^2}\right)\left(\sum_{i=1}^{L_S}\sigma^2(\widehat{\pi}_{Si}) + \sum_{i=1}^{L_S}\sum_{j=i+1}^{L_S}\sigma(\widehat{\pi}_{Si}, \widehat{\pi}_{Sj})\right). \tag{S4}$$

The double summation on the right, the sum of covariances of all of the site-specific estimates of π_{Si}, is a function of linkage disequilibrium; as noted in the main text, this is weak, but because there are $L_S(L_S - 1)$ covariance terms, it can be cumulatively nonneglible. A likely upper bound to the full term can be assigned by noting (from Figure 7 in the main text) that the correlation of zygosity among sites within genes in the study population is 0.04 for adjacent sites, and then progressively declines to 0.01 for sites separated by 100 bp and further to 0.005 for sites separated by 1000 bp. Full gene lengths (including introns) are typically well over 1000 bp in *D. pulex,* so we will take $0.01\sigma^2(\widehat{\pi}_{Si})$ to be the upper bound to each covariance term. This leads to a conservative estimator of the sampling variance,

$$\sigma^2(\widehat{\pi}_S) \simeq \left(\frac{1 + 0.01L_S}{L_S^2}\right)\left(\sum_{i=1}^{L_S}\sigma^2(\widehat{\pi}_{Si})\right). \tag{S5}$$

The standard error of $\widehat{\pi}_S$ is the square root of this expression, and an identical expression applies to replacement-site variance with N substituted for S in the subscripts.

Multiple substitutions per nucleotide site can cause Equation (S1) to underestimate the true level of divergence. For intra-population analyses, allelic divergence is generally low enough that the chances of such events is extremely small and can be ignored, and this is true for the vast majority of genes in this study. However, for divergences > 0.1, such effects start to become important. To guard against this problem, we utilized the Jukes-Cantor (1969) equation, which assumes an asymptotic divergence of 0.25, and is justified by the fact that nucleotide composition at silent and replacement sites in *D. pulex* (Table 1, main text) implies asymptotic divergences of 0.254 and 0.255 respectively. The corrected measure of the average number of substitutions per silent site is

$$\widetilde{\pi}_S = -(3/4)\ln[1 - (4\widehat{\pi}_S/3)]. \tag{S6}$$

From Ota and Nei (1994), the sampling variance of $\widetilde{\pi}_S$ is

$$\sigma^2(\widetilde{\pi}_S) \simeq \frac{\sigma^2(\widehat{\pi}_S)}{[1 - (4\widehat{\pi}_S/3)]^2}, \tag{S7}$$

again with an expression of the same form applying for replacement sites, and the standard error being the square root of Equation (S7).

Estimation of between-population divergence. To evaluate the levels of divergence for more distantly related alleles, we relied upon the two assembled genomes, one from the state of Oregon and the other from a sampling site near KAP. Working with each assembled haploid genome, and confining the analyses to the same types of codons noted above, the fractions of silent and replacement sites containing different nucleotides are d_S and d_N. To account for the possibility of multiple substitutions per site, these are then entered into the Jukes-Cantor equation to yield estimates of the average number of substitutions per site,

$$D_S = -(3/4)\ln[1 - (4d_S/3)]. \tag{S8}$$

The sampling variance of D_S is

$$\sigma^2(D_S) \simeq \frac{d_S(1-d_S)}{L_S \cdot [1-(4d_S/3)]^2}. \tag{S9}$$

Again, an expression of the same form applies for replacement sites, and the standard error of D_S is the square root of Equation (S9). No allowance is made here for nonindependence of divergence among sites because the isolates being compared are distant enough that sampling covariance associated with linkage-disequilibrium is expected to be minimal.

Measures of selection. Common measures of selection within a coding region are the ratios of replacement-site to silent-site diversity and divergence, i.e., $\widehat{\pi}_N/\widehat{\pi}_S$ and D_N/D_S. The standard errors of such ratios can be approximated by use of the general formula

$$\mathrm{SE}(u/v) = (u/v) \cdot \left(\frac{\sigma^2(u)}{u^2} + \frac{\sigma^2(v)}{v^2}\right)^{0.5}, \tag{S10}$$

where it is assumed that the numerator u and denominator v are statistically uncorrelated, and $u = \pi_N$ and $v = \pi_S$, or $u = D_N$ and $v = D_S$. Having obtained the standard errors of the two ratios of silent- and replacement-site divergence, the above formula can then be used to obtain the standard error of a common measure called the neutrality index,

$$\mathrm{NI} = \frac{\widetilde{\pi}_N/\widetilde{\pi}_S}{D_N/D_S}, \tag{S11}$$

letting $u = \widetilde{\pi}_N/\widetilde{\pi}_S$ and $v = D_N/D_S$ in Equation (S10). A related statistical indicator as to whether a sequence is evolving in a neutral manner at either level (e.g., $\pi_N = \pi_S$ or $D_N = D_S$) is given by the ratio of the difference to the expected standard error of the difference, e.g.,

$$T_\pi = \frac{\widetilde{\pi}_N - \widetilde{\pi}_S}{[\sigma^2(\widetilde{\pi}_N) + \sigma^2(\widetilde{\pi}_S)]^{0.5}}, \tag{S12}$$

with an analogous expression for the deviation between D_N and D_S.

Literature Cited

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules, pp. 21-132. In H. N. Munro (ed.), Mammalian Protein Metabolism. Academic Press, New York, NY.

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

Nei, M., and A. K. Roychoudhury. 1974. Sampling variances of heterozygosity and genetic distance. Genetics 76: 379-390.

Ota, T., and M. Nei. 1994. Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. Mol. Biol. Evol. 11: 613-619.