Patterns of genetic variability in genomic regions with low rates of recombination

Patterns of genetic variability in genomic regions with low rates of recombination

Hannes Becher[1], Benjamin C. Jackson and Brian Charlesworth

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Charlotte Auerbach Road, Edinburgh EH9 3FL

[1] Corresponding author and lead contact

Correspondence: hbecher@ed.ac.uk

SUMMARY

The amount of DNA sequence variability in a genomic region is often positively correlated with its rate of crossing over (CO) [1-3]. This pattern is caused by selection acting on linked sites, which reduces genetic variability and biases the frequency distribution of segregating variants towards more rare variants than are expected without selection (skew). These effects may involve the spread of beneficial mutations (selective sweeps, SSWs), the elimination of deleterious mutations (background selection, BGS) or both, and are expected to be stronger with lower CO rates [1-3]. However, in a recent study of human populations, the skew was reduced in the lowest CO regions compared with regions with somewhat higher CO rates [4]. A low skew in very low CO regions, compared with theoretical predictions, is seen in the population genomic studies of *Drosophila simulans* described here and in other *Drosophila* species. Here, we propose an explanation for lower than expected skew in low CO regions, and validate it using computer simulations; explanations for higher skew with higher CO rates, as in *D. simulans*, will be explored elsewhere. Partially recessive, linked deleterious mutations can increase neutral variability when the product of the effective population size (N_e) and the selection coefficient against homozygous carriers of mutations (s) is ≤ 1, i.e. there is associative overdominance (AOD) rather than BGS [5]. AOD can operate in low CO regions, producing a lower skew than in its absence. This opens up a new perspective on how selection affects patterns of variability at linked sites.

Results and Discussion

Diversity Statistics in Relation to CO Rates in *Drosophila simulans*

The top two panels of Figure 1 show the relations between the rate of crossing over and mean pairwise diversity at four-fold degenerate nucleotide sites (π_4), for two populations of *Drosophila simulans*. Consistent with previous studies of many species [1-3, 6], there is a significantly positive relation between CO rate and nucleotide site diversity for both X chromosome (X) and autosomes (A). As reported previously [7], X has much lower diversity than A for all bins of CO rates. Diversity is higher for the Madagascan than the Kenyan population, consistent with the latter having been founded as a relatively small population by flies descended from the putatively ancestral Madagascan population [7].

The bottom panels of Figure 1 show the relations between CO rate and a measure of the skew of the site frequency spectrum (SFS) towards rare variants, $\Delta\theta_{w4} = 1 - \pi_4 / \theta_{w4}$, where π_4 and θ_{w4} are the estimates of diversity based on the mean pairwise difference between alleles [8] and the number of segregating sites [9], respectively. Other measures of skew behave similarly (Figure S1). As was also found in a Rwandan population of *D. melanogaster* [10], there is much greater skew on X than A, although the absolute level of skew for both X and A is much higher than in *D. melanogaster*, suggesting a more intense recent population expansion in *D. simulans*. The skew is also larger in the Madagascan than the Kenyan population, consistent with the latter having experienced a bottleneck in population size or a slower rate of population expansion. The parameters of the population expansion for the Madagascan population were estimated by [11].

The most surprising feature of these results is that there is a nearly monotonic increase of $\Delta\theta_{w4}$ with CO rate in the autosomal CO regions. The overall autosomal values of $\Delta\theta_{w4}$ for non-CO regions are 0.22 and 0.13 for Madagascar and Kenya, respectively. A meta-analysis of 10 independent autosomal non-CO regions from six *Drosophila* species (Table S1) gave a mean of 0.181, with s.e. 0.036. The pattern is more complex for the CO regions of the X in *D. simulans* – there is a significant quadratic relation between $\Delta\theta_{w4}$ and CO rate, with an initial increase followed by a decrease at the highest CO rates. The non-CO regions of the X show values of $\Delta\theta_{w4}$ comparable to the CO regions, with the exception of the telomere in the Madagascan population, which has a value of 0.44, possibly reflecting a recent selective sweep, consistent with its very low π_4 value.

If the skew in the SFS were caused by selective sweeps (SSWs) alone, it would be expected to decrease, not increase, with CO rate, as was found in computer simulations with parameter values that are realistic for *Drosophila* [12]. If background selection (BGS) also contributes, these simulations showed that it should only produce a strong effect on skew in the non-CO regions, and cause these to have higher skews than regions with CO. Indeed, the simulations predicted higher skews in the non-CO regions than the observed autosomal means presented above: $\Delta\theta_{w4}$ was 0.24 for autosomal simulations with BGS and SSWs, and 0.21 with BGS alone, for a similar sample size to those in *D. simulans*. Given that demographic effects probably contribute substantially to the skew in the SFS in *D. simulans*, especially in Madagascar [11], an even higher skew in the non-CO regions is expected.

In this paper, we focus on the fact that the level of skew in the non-CO regions is smaller than expected with BGS and/or SSWs. We will address elsewhere the question of

relations between the level of skew of the SFS towards low frequency variants and the CO rate outside the non-CO regions, and simply note that it probably reflects the joint effects of SSWs and population size changes. These vary among genomic regions with different effective population sizes (N_e) associated with different CO rates, so that low N_e can reduce the effect on skew of a recent population expansion. This possibility was mentioned in a previous paper, in relation to the patterns of skew in a *D. melanogaster* population, where the skew increases at very high CO rates [10]; further modelling of the joint effects of demography and selection at linked sites is needed to test it rigorously.

Simulation Results and an Explanation for Low Skew with Low CO rates

The simulations in [12] assumed a dominance coefficient (h) of 0.5 for deleterious mutations, where the fitness of the heterozygous carriers of a deleterious mutation is reduced by hs relative to wild-type, with homozygotes suffering a reduction of s. However, studies of the effects of deleterious mutations on fitness components [13-15] suggest that there is likely to be partial recessivity of most slightly deleterious mutations, with $0 \leq h < 0.5$. Because selection against rare deleterious mutations in randomly mating populations acts mainly on their heterozygous carriers [16], the value of h does not greatly affect the fate of rare deleterious mutations with similar values of hs, so that $h = 0.5$ was used previously for convenience.

Empirical estimates of the parameters of the distribution of deleterious mutational effects on fitness (DFE) in *Drosophila* suggest that it has a large standard deviation relative to its mean [17-19]. Under a gamma distribution, the shape parameter for both nonsynonymous (NS) and UTR mutations has been estimated to be approximately 0.3 [20]. This raises the possibility that a subset of partially recessive deleterious mutations could cause associative overdominance (AOD), whereby loss of variability at neutral sites is retarded by linkage disequilibrium (LD) between selected and linked neutral sites [21]. Such LD causes a correlation in homozygosity between sites, so that homozygotes for variants at the neutral sites are associated with reduced fitness compared to heterozygotes when deleterious mutations have some degree of recessivity [5]. Recent theoretical work has shown that partially recessive deleterious mutations can result in AOD when $2N_es$ is less than 2.5 with sufficiently close linkage between neutral and selected sites [5], which is prevalent in low recombination genomic regions. Such a low value applies to a significant proportion of mutations when the DFE is sufficiently wide. For example, with $h = 0.2$, a shape parameter

of 0.3 and $2N_e s = 5000$ (the value for NS sites in our simulations), the proportion of mutations with $2N_e s \leq 2.5$ is 0.08. By definition, AOD acts similarly to true overdominance, a form of balancing selection. Long-term balancing selection is known to increase the lengths of the internal branches of coalescent trees relative to the total tree size [22-24]; AOD should have a similar effect, causing the SFS to become biased towards more frequent variants than with strict neutrality. BGS without AOD has the opposite effect, especially in low CO regions where selective interference among deleterious mutations increases the skew towards rare variants [25, 26].

If the more abundant strongly selected mutations cause standard BGS effects in a low recombination region, but some weakly selected mutations cause AOD, the net outcome for a low CO region should be lower diversity and a site frequency spectrum (SFS) more skewed towards low frequency variants than with no selection, but with higher diversity and less skew than with standard BGS. This process could account for the relatively low skew towards low frequency variants in the non-CO regions of humans [4] and six *Drosophila* species (Table S1).

We have investigated this possibility by simulations using a range of different dominance coefficients (see STAR Methods for details), and assuming a fixed population size. Figure 2A shows results with $h = 0.2$ and $h = 0.5$, for a range of CO rates that are all ≤ the mean rate of 2×10^{-8} for female meiosis in *D. melanogaster* [27, 28]. With deleterious mutations alone, advantageous mutations alone, and with both deleterious and advantageous mutations, the mean value of neutral diversity increases as the CO rate increases, whereas the degree of skew towards low frequency variants decreases. In the presence of deleterious mutations, the degree of skew in very low CO rate regions is substantially lower with $h = 0.2$ than $h = 0.5$, both in the presence (0.16, 0.24) and absence of advantageous mutations (0.10, 0.21), and is not far from the observed mean autosomal value across *Drosophila* species of 0.18. In addition, neutral diversity with deleterious mutations is higher for $h = 0.2$ (0.0041) than $h = 0.5$ (0.0024) in non-CO regions. Figure 2B shows the behaviour of a non-CO region with a range of dominance coefficients, showing that low h greatly reduces the skew caused by BGS, and increases diversity. As was found previously [12], BGS is the main cause of reduced variability in the non-CO cases, reflecting the fact that the rate of sweeps is greatly reduced by BGS when CO is absent, consistent with the empirical evidence that rates of adaptive evolution are greatly reduced in non-CO regions [12]. It is likely, therefore, that the effects of selection against deleterious mutations are the main cause of the low skew in these

regions. Our findings are consistent with the hypothesis that AOD (caused by the relatively small fraction of mutations that are weakly selected) reduces the effect of BGS when mutations are partially recessive.

Linkage Disequilibrium Patterns and a Further Test for AOD

This hypothesis can be further tested by noting that that AOD increases the relative lengths of internal branches of the coalescent tree for a neutral site (see above). By analogy with the effects on LD of a recent population size change, this change in tree topology increases the magnitude of LD [29, 30]. Conversely, BGS with low CO rates has the opposite effect, since it increases the relative lengths of the external branches of coalescent trees, as noted above. We can use the simulated values of neutral diversity to estimate the corresponding value of N_e from the formula $\pi = 4N_e u$, where u is the mutation rate per site) [31]. This can be used to determine the approximate expected value of r^2, the squared correlation coefficient between the allelic states of pairs of neutral sites, from equation (3) in [32]. As can be seen in Figure 3 (top left), the simulation values of median r^2 for all pairs of neutral and without CO (but with gene conversion) are more than an order of magnitude lower than the neutral expectations, which are 0.200 with $h = 0.2$ and 0.209 with $h = 0.5$, indicating a strong influence of selection at linked sites on tree shape. Nonetheless, the magnitude of LD with $h = 0.2$ is much greater than with $h = 0.5$ despite a higher level of diversity and hence N_e (note that low N_e is expected to increase r^2).

The bottom panels of Figure 3 show plots of median r^2 for 4-fold sites against physical location for 50 bins of distances between pairs of variants across genes on the non-CO chromosome four in the Madagascan and Kenyan populations. As is expected from bottleneck effects or a lower rate of population expansion, LD is markedly higher in Kenya than Madagascar, but the overall levels of LD are of similar magnitude to the simulation results, and are much lower than predicted from the estimates of N_e using π_4 and the *D. melanogaster* mutation rate [33]. There is an indication that there is a low level of crossing over on this chromosome, from the significant negative relation between r^2 and distance between pairs of sites, which is not observed in simulations with gene conversion alone.

It is, of course, possible that true balancing selection, rather than AOD, could contribute to these patterns. Figure S2 shows plots of 4-fold and 0-fold diversities per gene along chromosome four, for the Madagascan and Kenya populations. There is one gene, FBgn0268752, which stands out as having unusually high diversities at both types of site in

both populations. In the case of Madagascar, there are two intermediate-frequency nonsynonymous variants (at positions 40149 and 40648) (Table S2), whereas Kenya is segregating for nonsynonymous variants only at 40648, possibly because of loss of variability after a bottleneck. The polymorphisms at sites 40149 and 40648 in Madagascar are not in LD with each other, whereas there are strong associations between one of the variants at 40149 and synonymous variants to its left (Table S2). It is, therefore, possible that there is balancing selection acting at this locus, which appears to be unique to *D. simulans*. However, there is no evidence that the LD created by the 40149 variant extends outside this locus, so that it does not contribute to the general patterns described above.

Previous theoretical work on AOD due to deleterious, partially recessive mutations suggested that it is likely only to occur in small populations, where there is indeed evidence that the rate of loss of variability for molecular markers is slower than expected from the neutral value of N_e [5]. This is because AOD requires $N_e s$ to be of the order of 1, which is likely to be met by only a small minority of deleterious mutations in a large, randomly mating population. However, when N_e for neutral sites is greatly reduced by selection at linked sites, the extent of randomly generated LD between neutral and selected sites will be considerably enhanced, and LD is the driving force for AOD [5].

The simulation results shown in Figures 2 and 3 suggest that, with partially recessive mutations and the wide distribution of *s* values suggested by population genomic analyses of the effects of deleterious mutations, there is a noticeable effect of AOD on the SFS and extent of LD at neutral sites when CO rates are very low. This interpretation can be tested by comparing the results of simulations with a narrow versus wide range of selection coefficients for deleterious mutations, keeping the mean constant (Figure 4). With a narrow range of *s*, diversity, skew and the magnitude of LD are independent of *h*; with the wide range of selection coefficients, diversity and LD decrease with *h*, but skew increases. This is what is expected if AOD is influencing the behaviour of neutral variants in low CO regions.

The properties of genomes or genomic regions with low rates of genetic recombination are of considerable interest for a range of biological questions, including evolution and variation in bacteria [34, 35], asexual higher organisms [36, 37], organisms with high rates of self-fertilisation [38, 39], and Y and W chromosomes [40-42]. There is a general expectation that genetic diversity and molecular signatures of adaptive evolution should be greatly reduced in such systems, as a result of selection at linked sites. This reduction in adaptation is important for understanding phenomena such as the degeneration of Y chromosomes and the lack of evolutionary success of asexual and highly selfing species.

However, comparisons between different species with different modes of reproduction that influence the recombination rate are often made difficult by confounding factors, such as differences in the extent of colonisation events involving population size bottlenecks [38, 39]. The use of comparisons among genomic regions with different recombination rates largely removes this difficulty, and has revealed many of the expected results [2, 3, 6]. Here, we have shown that an unexpected pattern emerging from such comparisons suggests that the largely neglected process of associative overdominance may influence patterns of variability when recombination rates are very low. Because it requires partial recessivity of the fitness effects of deleterious mutations, and the opportunity for them to be present in both heterozygous and homozygous states, it will only operate in diploid or polyploid organisms with some degree of outbreeding. It will, therefore, not affect haploid, asexual or highly selfing organisms, or effectively haploid chromosomes such as Y and W. This suggests that we might expect to see differences in their patterns of variability compared with the non-CO regions of outbreeding diploid species.

Acknowledgements

We wish to thank Ching-Ho Chang and Amanda Larracuente for providing the sequence diversity data for the dot chromosome in *D. pseudoobscura* and *D. miranda*. We made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/). This research was funded by a Leverhulme Trust grant (IRPG-2015-033) to BC.

Author contributions

BC designed the project and contributed to the analyses of the data and simulations; HB conducted the simulations and contributed to the analyses of the population data and simulations; BCJ conducted the bioinformatic analyses of the raw sequencing data. All three authors contributed to the writing of the manuscript.

Declaration of Interests

The authors declare no competing interests.

Figure Legends

Figure 1. Genetic Diversity Statistics for Four-fold Degenerate Sites in Two Populations of *Drosophila simulans* in Relation to the Rate of Crossing Over. All panels share the same X axis, the mean proportion of the standard crossover (CO) rate per megabase in homologous regions of *Drosophila melanogaster* for each group of genes. Estimates of the pairwise nucleotide diversity (π_4) for each group are shown in the top panels, and estimates of a measure of the skew of the site-frequency spectrum ($\Delta\theta_{w4}$) in the bottom panels. These statistics were computed for each gene and then binned by CO rate, after weighting by the number of four-fold sites per gene in each bin. Autosomal bins are shown in black and X chromosomal bins in grey. The solid lines indicate significant regression fits for the X chromosomal values for the CO regions, with the following significance levels for the slope (π_4) or quadratic terms ($\Delta\theta_{w4}$) – Madagascar π_4: $p = 0.00960$; Madagascar $\Delta\theta_{w4}$: $p = 0.0083$; Kenya π_4: $p = 0.00598$; Kenya $\Delta\theta_{w4}$: $p = 0.00262$. The dashed lines indicate linear regression fits for the autosomal bins. The significance levels from Spearman rank correlations were as follows – Madagascar π_4: $\rho = 0.939$, $p < 2.2 \times 10^{-16}$; Madagascar $\Delta\theta_{w4}$: $\rho = 0.891$, $p = 0.00138$; Kenya π_4: $\rho = 0.927$, $p = 0.000130$; Kenya $\Delta\theta_{w4}$: $\rho = 0.685$, $p = 0.0351$). Non-CO regions are shown as filled symbols. These are spread out for better visibility and are labelled with their identities; Xc is the centromeric region of the X and Xt is its telomeric region; 2 and 3 are centromeric regions of chromosomes 2 and 3, respectively; 4 is the dot chromosome. See also Figure S1 for further measures of skew, Table S1 for a meta-analysis of skew for non-CO regions of 6 different *Drosophila* species, and Table S4 for the statistics for each bin.

Figure 2. Expectations for the Patterns of Genetic Diversity Statistics from Forwards-in-Time Simulations with Constant Population Size.

(A) The pairwise diversities at simulated autosomal synonymous sites (π, top row) and a measure of the distortion of the site-frequency spectrum ($\Delta\theta_w$, bottom row) for different rates of crossing over, expressed as a proportion of the mean sex-averaged rate for autosomes in *D. melanogaster*. Gene conversion was allowed with the standard parameter values for *D. melanogaster*, regardless of the CO rate. The theoretical expectations with no selection are given by the horizontal green lines. Simulations with selection against deleterious mutations alone are indicated as 'Deleterious', simulations with advantageous mutations alone as 'Advantageous', and those with both kinds of selection as 'Adv + del'. The results in the left-hand panels are for simulation runs with semi-dominant mutations ($h = 0.5$), and those in the right-hand panels are with partial recessivity ($h = 0.2$). Each data point shows the mean and 95% confidence intervals for 20 replicate simulations, for a group of 70 genes subject to selection (see STAR Methods for details of the parameters used in the simulations).
(B) Simulated values of π and $\Delta\theta_w$ for a range of different dominance coefficients (h) without crossing over but with gene conversion occurring at the standard rate for *D. melanogaster*.
See Table S1 for a comparison of levels of $\Delta\theta_{w4}$ between multiple species of *Drosophila*. For the detailed numerical outputs of the simulations, see Table S5.

Figure 3. Linkage Disequilibrium (LD) as Measured by the Median of r^2 Between Pairs of Sites The top panels show simulation results for a diploid population size of $N = 2500$ with selection against deleterious mutations and sweeps, with either no CO or CO at 1% of the standard rate. Black circles indicate runs without dominance ($h = 0.5$) and grey triangles indicate runs with partial recessivity ($h = 0.2$). Mantel tests carried out on individual simulation runs are only significant in the presence of CO. The bottom panels show the average values of the LD measure across bins of pairwise distances, between four-fold degenerate variant sites on chromosome four for the Madagascan and Kenyan populations. Mantel tests carried out on the (un-binned) data show that there is a significant decay of LD with increasing distance (Madagascar four-fold: 222 sites, Mantel statistic $R = 0.061$, $p = 0.001$; Madagascar zero-fold: 320 sites, $R = 0.040$, $p = 0.001$; Kenya four-fold: 187 sites, $R = 0.026$, $p = 0.015$; Kenya zero-fold: 267 sites, $R = 0.027$, $p = 0.002$).

Figure 4. Weakly Deleterious Mutations are Required for AOD Effects on Variability.

The panels show the results of simulations with deleterious mutations alone, in the absence of CO and a range of dominance coefficients, with the mean strength of selection being kept constant. Black and grey indicate two different distributions of deleterious fitness effects with the same mean but different shape parameters (β). The results with β = 0.3 (a wide distribution of selection coefficients), are the same as those shown previously, and are contrasted with the results for a narrow distribution (β = 10), where AOD is unlikely to occur.

STAR Methods

LEAD CONTACT AND MATERIALS AVAILABILITY

Requests for further information and for resources and reagents should be directed to, and will be answered by, the Lead Contact, Hannes Becher (hbecher@ed.ac.uk). This study did not generate new unique reagents.

METHOD DETAILS

Sampling, sequencing, and variant calling

The results described here are partly based on whole genome sequencing data available from the European Nucleotide Archive, study accession number: PRJEB7673. These data had been generated from 22 isofemale lines of *D. simulans*, established from fertilized females collected in Kenya and Madagascar. Eleven of these were collected by William Ballard in 2002 from Madagascar, and the other eleven were collected by Peter Andolfatto in 2006 from Kenya. These lines were maintained in the laboratory of P. Andolfatto for several generations, and subsequently inbred by brother-sister mating in the Charlesworth laboratory. The details of the breeding procedures and sequencing methods are fully described in [7].

Genomic DNA was prepared for each isofemale line by pooling 25 females, snap freezing them in liquid nitrogen, extracting DNA using a standard phenol-chloroform extraction protocol with ethanol, and ammonium acetate precipitation. These flies were sequenced by the Beijing Genomics Institute (BGI; http:// bgi-international.com/). A 500-bp short-insert library was constructed for each sample, and the final data provided consisted of

90-bp paired-end Illumina sequencing (pipeline version 1.5), with an average coverage of 64X.

We also obtained sequence data on 20 further *D. simulans* isofemale lines from the European Nucleotide Archive, study accession number: PRJNA215932. These lines were from the same sampling localities in Kenya and Madagascar as above. Each line had been sequenced on between 2-3 lanes of paired-end Illumina sequencing at the UC Irvine High Throughput Genomics centre (http://ghtf.biochem.uci.edu/) per line. For further information, see [7].

We downloaded the raw read data in fastq format from the European Nucleotide Archive. We mapped these reads to version 2.02 of the *D. simulans* genome (FlyBase release 2017_04) [43] using BWA MEM [44], then sorted, merged and marked duplicates on the resulting BAM files using Picard Tools version 2.8.3 (https://broadinstitute.github.io/picard/). We called variants separately for each individual line using the HaplotypeCaller tool from GATK version 3.7 [45] with the options –emitRefConfidence, BP_RESOLUTION and –max-alternate-alleles 2, then made per-chromosome VCF files for the whole population using the GATK v3.7 tools combineGVCFs and genotypeGVCFs. All the scripts necessary for downloading the fastq files and calling variants are available at https://github.com/benja-mincjackson/dsim_variant_pipeline_ref_v2.02.git. We defined sites as 4-fold degenerate in all transcripts using information from the gff format annotation of the *D. simulans* genome v2.02 (available from ftp://ftp.flybase.net/genomes/Drosophila_simulans/dsim_r2.02_FB2017_04/ gff/dsim-all-r2.02.gff.gz). As described in [7], we excluded individuals with high residual heterozygosity and with unusual patterns of pairwise synonymous diversity between the populations, leaving us with sequences from 18 Kenyan and 21 Madagascan lines.

Data analyses

In the absence of a *D. simulans* map based directly on analyses of the products of meiosis (but see [46] for an LD map from a Portugese population), we used the mapping data of [27] for *D. melanogaster*. The *D. melanogaster* crossing over (CO) rates were smoothed by Loess regressions against the physical positions of markers in the genome. Each *D. simulans* gene was assigned the CO rate per megabase in female meiosis of the corresponding position in *D. melanogaster,* using an alignment of the genome sequences of *D. simulans* (v2.02) and *D. melanogaster* (v5.57) [43]. Following [10], the absence of recombinational exchange between homologous chromosomes in *Drosophila* males was taken into account by

multiplying the autosomal CO rates by one-half, and the X linked rates by two-thirds. All CO rates presented below have been corrected in this way.

For genomic regions with non-zero CO rates, genes were grouped into ten equally-sized bins. For the autosomes (A), there were approximately 730 genes per bin, and approximately 180 genes per bin for the X chromosome (X). All *D. simulans* genes that were aligned to *D. melanogaster* scaffold heterochromatin, or to regions defined as lacking CO in [10], were excluded before binning. These non-CO genes were then binned by chromosome (2: 86, 3: 54, X (tip): 23, X (centromere): 33), and were used for the analyses of the non-CO regions described below.

Chromosome arm 3R, which contains a large inversion relative to *D. melanogaster* [47], was excluded from our analyses. The genetic maps of the X chromosome appear to be similar in the two species [48] so that we used the *D. melanogaster* CO rate estimates as a proxy for those in *D. simulans*. For the analyses of chromosome 2 and 3L, the binning procedure assumes the rank order of the CO rate of a gene to be related to its physical position in the same way as in *D. melanogaster*, even though the absolute CO rates per megabase differ between the species: *D. simulans* has higher CO rates, especially in the low CO rate regions near the centromere and telomere [48]. Non-parametric rank correlations were therefore used to assess statistical relationships between the summary statistics for the autosomal polymorphism data and CO rates.

Levels of polymorphisms per gene were measured by estimates of mean pairwise nucleotide site diversity (π) [8] and Watterson's theta (θ_w) [9] for fourfold and zero-fold sites, distinguished here by subscripts 4 and 0 (see Figure S2). Nucleotide positions with missing data (no calls for one or more individuals) were excluded. The extent of skew of the folded site frequency spectra (SFS) at fourfold and zero-fold sites for a CO bin was assessed in various ways: the proportion of singleton variants (P_S), the mean value of Tajima's D statistic (D_T) across genes in a bin [49], and the relative difference between the mean values of θ_{w4} and π_4 ($\Delta\theta_{w4} = 1 - \pi_4 / \theta_{w4}$) for the genes in a bin. This is related to the $\Delta\pi$ statistic of [50], but has the opposite sign and is not multiplied by Watterson's correction factor for sample size [9]. An excess of rare variants over the equilibrium neutral expectation for mutation and drift under the infinite sites model [31] is indicated by an excess of P_S over the theoretical value of the probability of observing derived variants present at frequencies of $1/n$ and $1 - 1/n$ in a sample of n alleles [9], by a negative value of D_T, and by a positive value of $\Delta\theta_w$. The last statistic has the advantage of being less dependent on the sample size than the

other statistics, and is used in the figures presented in the text. Results for the other statistics are shown in Figure S1.

LD was computed between pairs of 4-fold sites as the square of Pearson's correlation coefficient, *r,* between the allelic states of the sites [51]. Pairs of sites were then grouped by distance into bins of 10 kbp and the mean r^2 was computed for each bin, discarding bins over 5×10^5 bp apart, since these contained very few datapoints. The haplotypes at the high-diversity site on the 4th chromosome are shown in Table S2.

Simulations

Forwards-in-time simulations were carried out with SLiM v2.6 [52]. The values of the deterministic parameters of mutation, selection and recombination were chosen on the basis of estimates from *D. melanogaster*, as described by [12, 20]. For the simulations, these values were multiplied by $(1.33 \times 10^6)/N$, where N is the number of diploid individuals used in the simulations, to ensure that the products of N and these parameters correspond approximately to the values for a real *Drosophila* population [12]; this rescaling should conserve the properties of the evolutionary process, with some exceptions discussed in [12]. In particular, rescaling affects the distribution of CO events if there is CO interference. This is not of concern here where low-CO regions are studied. Population sizes of 2500, 10000 and 25000 diploid individuals were simulated, with a 1:1 sex ratio. The genome consisted of 69 intergenic regions of 4 kb each, and 70 genes plus 4kb of non-coding sequence at both ends of the genome. Each gene was made up of a 5′ and a 3′ UTR of 190bp and 280bp, respectively, and five coding exons of 300 bp separated by four 100bp introns, resulting in a total gene length of 2370 bp and a total chromosome length of 449900bp. This provides an approximate model of the *Drosophila* fourth chromosome, the best-studied non-CO genomic region [53].

A constant unscaled per-base pair mutation rate of 4.5×10^{-9} was assumed. To speed up the simulations, no mutations were allowed in the intergenic regions, but CO and gene conversion were allowed at uniform rates over the whole region. Following earlier practice [12], all intronic mutations were neutral. For simulations that included both background selection (BGS) and selective sweeps (SSWs), UTR mutations were either deleterious or beneficial, with a probability of being beneficial of $p_u = 9.04 \times 10^{-4}$ (see Table S3). Coding regions were assumed to consist of 30% 4-fold degenerate sites and 70% 0-fold sites. 30% of new mutations were thus neutral; of the remainder, most were deleterious, with a probability

of $p_a = 2.21 \times 10^{-4}$ that a mutation at such a site was advantageous [12]. In addition, simulations were run with SSWs only, or with BGS only. Some purely neutral simulations were also run for the purpose of testing. For these configurations, the appropriate mutation types were replaced by neutral ones.

Gamma distributions of selection coefficients were used for the deleterious mutations, and exponential distributions for the beneficial mutations (see Table S3). Since the main question investigated here is the possibility of associative overdominance effects in genomic regions with low CO rates, which requires partial recessivity of deleterious mutations [5], we adjusted the scaled selection coefficients shown in Table S3 by a factor of $1/(2h)$ for runs with dominance coefficient h that differed from 0.5, in order to maintain the same heterozygous fitness effects of mutations. In large, randomly mating populations of the type used here, the fates of new mutations are largely determined by their heterozygous fitness effects [16, 54]. The value of h should thus not greatly affect the rates of fixation of deleterious and beneficial mutations, or their probability distributions at segregating sites, provided that it is sufficiently far from zero.

Following [12], the standard CO rate in female meiosis was set to 2×10^{-8}, and a range of values below this, including no CO, was simulated. The rate of initiation of gene conversion (GC) tracts in female meiosis was set to 4×10^{-8}, double the observed rate for *D. melanogaster* [28], to compensate for the fact that gene conversion in SLiM v2.6 assumes that tracts are initiated in only one direction. The mean tract length was 440 bp, and individual values were drawn from a geometric distribution. No recombination was allowed in males, so that the evolutionarily effective autosomal recombination rate is half the assigned rate.

The simulations were run for 35,000 (= 14N) generations and a sample of 20 genomes was taken every N generations. The first 20,000 (= 8N) generations were treated as a burn-in period, during which the population approached mutation-selection-drift equilibrium, except for runs with zero crossing over, where equilibrium was approached very quickly because of strong effects of selection at linked sites. A few simulations were carried out with larger values of N (10,000 and 25,000) in the absence of crossing over. For these, we used the same burn-in and run times, because the simulations reached equilibrium much sooner than with crossing over. Summary statistics were calculated from the state of the population after the simulation finished, and fixations of new mutations were recorded if they arose after the burn-in period.

The large population sizes were used to check whether the rather strong selection against deleterious mutations with a population size of 2,500 caused a very strong haplotype structure in the absence of CO. This can occur because of the build-up of pseudo-overdominance between complementary haplotypes of the type 1010… /0101… , where 1 denotes the wild-type state at a site, and 0 denotes the mutant state [55-58]. However, there were relatively small differences in outcomes for the different population sizes, and even the smallest population size showed no evidence for such an effect (see Table S5).

QUANTIFICATION AND STATISTICAL ANALYSIS

For the analysis of the *D. simulans* population genetic statistics, genes were binned by crossover rate (10 bins of approx. 730 genes for autosomes, 10 bins of approximately 180 genes for the X chromosome) as detailed in Table S4. Per-bin means of π_{4w} and θ_{4w}, weighted by the number of relevant sites per gene, were analysed with Spearman rank correlations for the autosomes, and with least-squares regression models (containing a quadratic term) for the X. Both methods were implemented in R. The *p* values reported are the overall values for the rank correlations, and for the quadratic terms in the X models.

The confidence intervals for the empirical and simulated data were generated via bootstrapping using the R package “boot”. For the empirical data, bootstrapping was carried out on the level of genes per bin (see Table S4). All simulations were run with 20 replicates, which provide the basis for bootstrapping. We report the 95% confidence intervals of the mean based on 1000 bootstrap replicates. For analyses of linkage disequilibrium, we used the same approach for the medians of the r^2 statistic described above.

The decay of linkage disequilibrium with pairwise distance between sites was assessed using Mantel tests as implemented in the R package “vegan”. The value of the Mantel statistic and the associated p-values are reported in the legend of Figure 3.

DATA AND CODE AVAILABILITY

The original data are available from the European Nucleotide Archive, accession numbers PRJEB7673 and PRJNA215932. Other datasets and code developed here are available on Zenodo, https://doi.org/10.5281/zenodo.3403084.

Titles of the Supplemental Tables

Table S4. Binned population genetic statistics from two populations of *Drosophila simulans*. Related to Figure 1 and Figure S1.

Table S5. Results of forwards-in-time simulations with parameters relevant to *Drosophila* populations. Related to Figure 2.

REFERENCES

1. Begun, D., and Aquadro, C.F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rate in *Drosophila melanogaster*. Nature *356*, 519-520.
2. Cutter, A.D., and Payseur, B.A. (2013). Genomic signatures of selection at linked sites: unifying the disparity among species. Nature Rev. Genet. *14*, 262-272.
3. Charlesworth, B., and Campos, J.L. (2014). The relations between recombination rate and patterns of molecular evolution and variation in *Drosophila*. Ann. Rev. Genet. *48*, 383-403.
4. Pouyet, F., Aeschbacher, S., Thiery, A., and Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demograpic inferences. eLife *7*, e36317.
5. Zhao, L., and Charlesworth, B. (2016). Resolving the conflict between associativeoverdominance and background selection. Genetics *203*, 1315-1334.
6. Corbett-Detig, R.B., Hartl, D.L., and Sackton, T.B. (2015). Natural selection constrains neutral diversity across a wide range of species. PloS Biology *13*, e1002112.
7. Jackson, B.C., Campos, J.L., Haddrill, P.R., Charlesworth, B., and Zeng, K. (2017). Variation in the intensity of selection on codon bias over time causes contrasting patterns of base composition evolution in *Drosophila*. Genome .Biol. Evol. *9*, 102-123.
8. Nei, M., and Tajima, F. (1983). DNA polymorphism detectable by restriction endonucleases. Genetics *97*, 145-163.
9. Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. *7*, 256-276.
10. Campos, J.L., Halligan, D.L., Haddrill, P.R., and Charlesworth, B. (2014). The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. Mol. Biol. Evol. *31*, 1010-1028.
11. Zeng, K., Jackson, B.C., and Barton, H.J. (2019). Methods for estimating demography and detecting between-locus differences in the effective population size and mutation rate. Mol. Biol. Evol. *36*, 423-433.
12. Campos, J.L., and Charlesworth, B. (2019). The effects on neutral variability of recurrent selective sweeps and background selection. Genetics *212*, 287-303.
13. Crow, J.F. (1993). Mutation, mean fitness, and genetic load. Oxf. Surv. Evol. Biol. *9*, 3-42.

14. Manna, F., Martin, G., and Lenormand, T. (2011). Fitness landscapes: An alternative theory for the dominance of mutation. Genetics *189*, 923-937.
15. Charlesworth, B. (2015). Causes of natural variation in fitness: evidence from studies of Drosophila populations. Proc. Natl. Acad. Sci. USA *12*, 1662-1669.
16. Haldane, J.B.S. (1927). A mathematical theory of natural and artificial selection. Part V. Selection and mutation. Proc. Camb. Phil. Soc. *23*, 838-844.
17. Loewe, L., Charlesworth, B., Bartolomé, C., and Nöel, V. (2006). Estimating selection on nonsynonymous mutations. Genetics *172*, 1079-1092.
18. Keightley, P.D., and Eyre-Walker, A. (2007). Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics *177*, 2251-2261.
19. Kousathanas, A., and Keightley, P.D. (2013). A comparison of models to infer the distribution of fitness effects of new mutations. Genetics *193*, 1197-1208.
20. Campos, J.C., Zhao, L., and Charlesworth, B. (2017). Estimating the parameters of background selection and selective sweeps in *Drosophila* in the presence of gene conversion. Proc. Natl. Acad. Sci. USA *114*, E4762-E4771.
21. Frydenberg, O. (1963). Population studies of a lethal mutant in *Drosophila melanogaster.* I. Behaviour in populations with discrete generations. Hereditas *50*, 89-116.
22. Nordborg, M., and Innan, H. (2003). The genealogy of sequences containing multiple sites subject to strong selection in a subdivided population. Genetics *163*, 1201-1213.
23. Zeng, K., Fu, X.-Y., Shi, S., and Wu, C.-I. (2006). Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics *174*, 1431-1439.
24. Charlesworth, B., and Charlesworth, D. (2010). Elements of Evolutionary Genetics, (Greenwood Village, CO: Roberts and Company).
25. Comeron, J.M., and Kreitman, M. (2002). Population, evolutionary and genomic consequences of interference selection. Genetics *161*, 389-410.
26. Gordo, I., Navarro, A., and Charlesworth, B. (2002). Muller's ratchet and the pattern of variation at a neutral locus. Genetics *161*, 835-848.
27. Comeron, J., Ratnappan, R., and Bailin, S. (2012). The many landscapes of recombination in *Drosophila melanogaster*. PLos Genetics *8*, e1002905.
28. Miller, D.E., Smith, C.B., Kazemi, N.Y., Cockrell, A.J., Arvanitakas, A.V., Blumenstiel, J.P., Jaspersen, S.L., and Hawley, R.S. (2016). Whole-genome analysis of individual meiotic events in *Drosophila melanogaster* reveals that noncrossover gene conversions are insensitive to interference and the centromere effect. Genetics *203*, 159-171.
29. Slatkin, M. (1994). Linkage disequilibrium in stable and growing populations. Genetics *137*, 331-336.
30. McVean, G.A.T. (2002). A genealogical interpretation of linkage disequilibrium. Genetics *162*, 987-991.
31. Kimura, M. (1971). Theoretical foundations of population genetics at the molecular level. Theor. Pop. Biol. *2*, 174-208.
32. Weir, B.S., and Hill, W.G. (1986). Nonuniform recombination within the human b-globin gene cluster. Am. J. Hum. Genet. *38*, 776-778.

33. Schrider, D.R., Houle, D., Lynch, M., and Hahn, M.W. (2013). Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. Genetics *194*, 937-954.
34. Neher, R.A. (2013). Genetic draft, selective interference and population genetics of rapid adaptation. Ann. Rev. Ecol. Evol. Syst. *44*, 195-215.
35. Price, M.N., and Arkin, A.P. (2015). Weakly deleterious mutations and low rates of recombination limit the impact of natural selection on bacterial genomes. mBio *6*, e01302-01315.
36. Agrawal, A.F., and Hartfield, M. (2916). Coalescence with background and balancing selection in systems with bi- and uniparental reproduction: Contrasting partial aexuality and selfing. Genetics *202*, 313-326.
37. Bast, J., Parker, D.J., Dumas, Z., Jalvingh, K.M., Van, P.T., Jaron, K.S., Figuet, E., Brandt, A., Galtier, N., and Schwander, T. (2018). Consequences of asexuality in natural populations: Insights from stick insects. Mol. Biol. Evol. *35*, 1668-1677.
38. Charlesworth, D., and Wright, S.I. (2001). Breeding systems and genome evolution. Curr. Opn. Genet. Dev. *11*, 685-690.
39. Wright, S.I., Kalisz, S., and Slotte, T. (2013). Evolutionary consequences of self-fertilization in plants. Proc. R. Soc. B. *280*, 20130133.
40. Charlesworth, B., and Charlesworth, D. (2000). The degeneration of Y chromosomes. Phil. Trans. R. Soc. B. *355*, 1563-2572.
41. Bachtrog, D. (2008). The temporal dynamics of processes underlying Y chromosome degeneration. Genetics *179*, 1513-1525.
42. Kaiser, V.B., and Charlesworth, B. (2010). Muller's ratchet and the degeneration of the *Drosophila miranda* neo-Y chromosome. Genetics *185*, 339-348.
43. Thurmond, J., Goodman, J.L., Strelets, V.B., Attrill, H., Gramates, L.S., Marygold, S.J., Matthews, B.B., Millburn, M., Antonazzo, G., Trovisco, V., et al. (2019). FlyBase 2.0: the next generation. Nucl. Acids Res. *47*, D759-D756.
44. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv *1303.3997*.
45. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., M., D., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next- generation DNA sequencing data. Genome Res. *20*, 1297-1303.
46. Howie, J.M., Mazzuco, R., Taus, T., Nolte, V., and Schlötterer, C. (2019). DNA motifs are not general predictors of recombination in two *Drosophila* sister species. Genome Biol. Evol. *11*, 1345-1357.
47. Sturtevant, A.H. (1921). A case of rearrangement of genes in Drosophila. Proc. Natl. Acad. Sci. USA *7*, 235-237.
48. Sturtevant, A.H. (1929). The genetics of *Drosophila simulans*. Carnegie Inst. Wash. Publ. *399*, 1-62.
49. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis. Genetics *123*, 585-595.
50. Langley, S.A., Karpen, G.H., and Langley, C.H. (2014). Nucleosomes shape DNA polymorphism and divergence. PloS Genetics *10*, e1004457.
51. Hill, W.G., and Robertson, A. (1968). Linkage disequilibrium in finite populations. Theor. Appl. Genet. *38*, 226-231.

52. Haller, B.C., and Messer, P.W. (2016). SLiM 2: Flexible, interactive forward genetic simulations. Mol .Biol. Evol. *34*, 230-240.
53. Riddle, N.C., and Elgin, S.C. (2018). The *Drosophila* dot chromosome: Where genes flourish amidst repeats. Genetics *210*, 757-772.
54. Fisher, R.A. (1922). On the dominance ratio. Proc. Roy. Soc. Edinburgh *42*, 321-341.
55. Charlesworth, D., Morgan, M.T., and Charlesworth, B. (1993). Mutation accumulation in finite outbreeding and inbreeding populations. Genet. Res. *61*, 39-56.
56. Pamilo, P., and Palsson, S. (1998). Associative overdominance, heterozygosity and fitness. Heredity *81*, 381-389.
57. Palsson, S., and Pamilo, P. (1999). The effects of deleterious mutations on linked neutral variation in small populations. Genetics *153*, 475-483.
58. Palsson, S. (2001). The effects of deleterious mutations in cyclically parthenogetic organisms. J. Theor. Biol. *208*, 201-214.