Background Selection as Baseline for Nucleotide Variation across the Drosophila Genome

Josep M. Comeron

University of Iowa

Copyright: © 2014 Josep M. Comeron


Hosted by Iowa Research Online. For more information please contact: lib-ir@uiowa.edu.
Background Selection as Baseline for Nucleotide Variation across the *Drosophila* Genome

Josep M. Comeron

1 Department of Biology, University of Iowa, Iowa City, Iowa, United States of America, 2 Interdisciplinary Program in Genetics, University of Iowa, Iowa City, Iowa, United States of America

### Abstract

The constant removal of deleterious mutations by natural selection causes a reduction in neutral diversity and efficacy of selection at genetically linked sites (a process called Background Selection, BGS). Population genetic studies, however, often ignore BGS effects when investigating demographic events or the presence of other types of selection. To obtain a more realistic evolutionary expectation that incorporates the unavoidable consequences of deleterious mutations, we generated high-resolution landscapes of variation across the *Drosophila melanogaster* genome under a BGS scenario independent of polymorphism data. We find that BGS plays a significant role in shaping levels of variation across the entire genome, including long introns and intergenic regions distant from annotated genes. We also find that a very large percentage of the observed variation in diversity across autosomes can be explained by BGS alone, up to 70% across individual chromosome arms at 100-kb scale, thus indicating that BGS predictions can be used as baseline to infer additional types of selection and demographic events. This approach allows detecting several outlier regions with signal of recent adaptive events and selective sweeps. The use of a BGS baseline, however, is particularly appropriate to investigate the presence of balancing selection and our study exposes numerous genomic regions with the predicted signature of higher polymorphism than expected when a BGS context is taken into account. Importantly, we show that these conclusions are robust to the mutation and selection parameters of the BGS model. Finally, analyses of protein evolution together with previous comparisons of genetic maps between *Drosophila* species, suggest temporally variable recombination landscapes and, thus, local BGS effects that may differ between extant and past phases. Because genome-wide BGS and temporal changes in linkage effects can skew approaches to estimate demographic and selective events, future analyses should incorporate BGS predictions and capture local recombination variation across genomes and along lineages.

### Citation


###-funded by US National Institutes of Health grant R22GM092501 and National Science Foundation grant DEB1354921. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Competing Interests

The author has declared that no competing interests exist.

* Email: josep-comeron@uiowa.edu

### Introduction

The causes of the variation observed within natural populations have been a long-standing question in evolutionary and genetic studies. Particular insight into these causes can be gained by analyzing the distribution of nucleotide diversity across genomes, where species- and population-specific parameters such as the number of individuals, environmental factors, or demography are constant. A number of population genetics models have been put forward to explain this intra-genomic variation in diversity, often including the predicted consequences that selection acting at a genomic site impinges on genetically linked sites, either neutral or under selection themselves (i.e., models of 'selection at linked sites'; [1–4] and references therein). Although there is general agreement that selection at linked sites can play a role shaping levels of variation, there is still intense debate and research on the selective nature of the mutations causing such effects (e.g., beneficial or deleterious) and whether the same causes can be applied to different species [4–6].

Strongly beneficial mutations rise rapidly to fixation and hitchhike adjacent sites, causing a fingerprint of reduced intra-specific variation around the selected site known as a 'selective sweep' (the HHss model; [2,3,7–11]). A qualitatively similar outcome can be generated by another model of selection and hitchhiking effects at genetically linked sites without requiring adaptive changes, just as a result of the continuous input of strongly deleterious mutations and their removal by natural selection (the background selection (BGS) model [1,4,12–16]). Both models also predict that the consequences of selection removing adjacent diversity diminish when genetic recombination increases, a general pattern that has been observed in many species when comparing genomic regions with high and low (or zero) recombination rates (reviewed in [1,3–5]).

The magnitude and distribution of recombination rates across genomes play key roles in predicting the consequences of selection on adjacent variation. In humans, for instance, the presence of large recombination cold spots raised the possibility that BGS could reduce polymorphism levels at specific genomic regions. In agreement, recent analyses using models of purifying selection rather than purely neutral ones suggest that patterns of nucleotide diversity across the human genome are consistent with BGS predictions [17–21]. In the model system *D. melanogaster*, low-resolution recombination maps described limited or absent recombination near sub-telomeric and –centromeric regions.
Author Summary

The removal of deleterious mutations from natural populations has potential consequences on patterns of variation across genomes. Population genetic analyses, however, often assume that such effects are negligible across recombinating regions of species like Drosophila. We use simple models of purifying selection and current knowledge of recombination rates and gene distribution across the genome to obtain a baseline of variation predicted by the constant input and removal of deleterious mutations. We find that purifying selection alone can explain a major fraction of the observed variance in nucleotide diversity across the genome. The use of a baseline of variation predicted by linkage to deleterious mutations as null expectation exposes genomic regions under other selective regimes, including more regions showing the signature of balancing selection than would be evident when using traditional approaches. Our study also indicates that most, if not all, nucleotides across the D. melanogaster genome are significantly influenced by the removal of deleterious mutations, even when located in the middle of highly recombinating regions and distant from genes. Additionally, the study of rates of protein evolution confirms previous analyses suggesting that the recombination landscape across the genome has changed in the recent history of D. melanogaster. All these reported factors can skew current analyses designed to capture demographic events or estimate the strength and frequency of adaptive mutations, and illustrate the need for new and more realistic theoretical and modeling approaches to study naturally occurring genetic variation.

whereas recombination outside these sub-telomeric and centromeric regions (i.e., across trimmed chromosome arms) has been often assumed to be both high and homogeneously distributed. As a consequence, variation in nucleotide diversity across trimmed chromosome arms has been mostly attributed to positive selection and selective sweeps ([4,5,22–29]; but see [30]).

There are, however, several reasons to believe that BGS effects could be significant in D. melanogaster as well. First, compared with humans, D. melanogaster has a more compact genome and a larger effective population size (Ne), predicting tighter genetic linkage between genes and stronger purifying selection, respectively, and both factors forecast greater BGS effects. Second, recent whole-genome studies of recombination rates in D. melanogaster exposed extensive heterogeneity in the distribution of crossover rates even after removing sub-telomeric and centromeric regions [31]. This high degree of variation in recombination rates across D. melanogaster chromosomes is observed when recombination is obtained from a single cross of two specific strains [31,32] as well as when analyzing a species’ average obtained from combining genetic maps from crosses of several natural strains [31]. The presence of coldspots of recombination embedded in chromosomal regions assumed to have high recombination rates, therefore, provides the opportunity for BGS to play a more significant role across broader genomic regions than previously anticipated [31]. Finally, Charlesworth [33] has recently showed that BGS effects are predicted to be detectable in the middle of recombining chromosome arms in D. melanogaster.

The consequences of BGS at a given nucleotide position in the genome (focal point) can be described by the predicted level of neutral nucleotide diversity when selection at linked sites is allowed (π) relative to the level of diversity under complete neutrality and free recombination between sites (π0), with \( B = \pi/\pi_0 \) [12–15]. Therefore, \( B \approx 1 \) would indicate negligible BGS effects whereas \( B \ll 1 \) would suggest very strong BGS and a substantial reduction in levels of neutral diversity. \( B \) can also be understood in terms of a reduction in Ne and variation in B forecasts differences in levels of diversity within species but also differences in the efficacy of selection, which can be approximated by the product of Ne and the selection coefficient s. Note, however, that the prediction about reduced efficacy of selection is a qualitative one since there is no simple scalar transformation of Ne influenced by selection at linked sites that allows estimating probabilities of fixation of selected mutations [34,35]. Thus, a comprehensive study of the predictive power of BGS to explain natural variation across genomes needs to show that, 1) conditions exists across a genome to generate significant overall effects reducing \( B \); 2) \( B \) varies across the genome, and 3) regions with reduced \( B \) are associated with reduced levels of polymorphism and efficacy of selection (e.g., detectable on rates of protein evolution).

Here, we investigated what is the fraction of the D. melanogaster genome that is influenced by BGS and how much of the observed variance in patterns of intra-specific variation and rates of evolution across this genome can be explained by BGS alone. Importantly, to obtain a sensible BGS baseline that could be used to test for positive selection and other departures from neutrality, we investigated BGS models that are purposely simple and independent of nucleotide variation data. Additionally, we studied whether our conclusions are sensitive to parameters of the BGS model. To this end, we expanded approaches previously applied to investigate human diversity [17,18] to now estimate BGS effects across the D. melanogaster genome.

In all, we generated a detailed description of the consequences of purifying selection on linked sites at every 1 kb along D. melanogaster chromosomes under a variety of BGS models. Our results show that BGS likely plays a detectable role across the entire genome and that purifying selection alone can explain a very large fraction of the observed patterns of nucleotide diversity in this species. Notably, we show that these conclusions are robust to different parameters in the BGS models. The use of a BGS baseline also uncovers the presence of regions with the signature of a recent selective sweep and, less expected, numerous instances of balancing selection. Furthermore, analyses of rates of protein evolution suggest that the recombination landscape has changed recently along the D. melanogaster lineage thus generating disparity between short- and long-term Ne at many genomic positions. We discuss the advantages of incorporating BGS predictions across chromosomes and the potential consequences of temporal variation in recombination landscapes when estimating demographic and selective events.

Results

BGS models

BGS expectations (i.e., estimates of \( B \)) were obtained for every 1-kb region across the whole genome as the cumulative effects caused by deleterious mutations at any other site along the same chromosome (see Materials and Methods for details). These estimates of \( B \) were based on BGS models that include our current knowledge of genome annotation at every nucleotide site of the genome and high-resolution recombination landscapes in D. melanogaster that distinguish between crossover and gene conversion rates [31]. These models also incorporate the possibility that strongly deleterious mutations occur at sites that alter amino acid composition as well as at a fraction of sites in noncoding sequences. The inclusion of deleterious mutations in noncoding sequences allows taking into account the existence of regulatory
and other non-translated functional sequences, either in introns and 5'- and 3'-flanking UTRs, or in intergenic regions [22,33,36–38]. For each category of selected sites (nonsynonymous, intronic, UTR, or intergenic) we used the proportion of constrained sites (8) estimated for D. melanogaster [22,37,30] as the fraction of sites with deleterious fitness consequences when mutated [33]. In terms of recombination rates, we studied BGS predictions following the standard approach of including crossover as the sole source of recombination (hereafter models MG,CO) and also when combining the effects of crossover and gene conversion events (models MCO,GC) to better quantify the true degree of linkage between sites in natural populations.

The distribution of deleterious fitness effects (DDFE) and the diploid rate of deleterious mutations per generation (U) are parameters that have direct implications on estimates of B but are more difficult to establish experimentally. Although a gamma distribution has been proposed a number of times for deleterious mutations [39–44], a log-normal DDFE allows capturing the existence of lethal mutations and fits better D. melanogaster polymorphism data [45,46]. Additionally, a log-normal DDFE predicts a higher fraction of mutations with minimal consequences removing linked variation than a gamma DDFE and, ultimately, weaker BGS effects (see Materials and Methods). Therefore, the use of a log-normal DDFE can be taken as a conservative approach when inferring the magnitude of BGS effects.

Direct estimations of deleterious mutation rates are still fairly limited. In D. melanogaster, initial analyses of mutation accumulation lines estimated a mutation rate for point mutations and small indels [u] of −8.4×10−9/bp/generation and a diploid rate of deleterious mutations per generation (U) of ~1.2 [47]. Nevertheless, one of the lines used in this study had an unusually high mutation rate [48] and more recent studies suggest u=4–5×10−9 (U~0.6) for point mutations and small indels [48–50]. These lower estimates, however, do not include the possible presence in natural populations of genotypes with high mutation rates or the deleterious consequences of transposable element (TE) insertions.

In fact, TE insertions are very abundant in natural populations of D. melanogaster [51–60] and have been proposed to be an important source of BGS in this species [30]. Therefore, U~0.6 represents a lower bound for the deleterious mutation rate when inferring the consequences of BGS. To include the consequences of TE insertion in our BGS models, we obtained an approximate diploid insertion rate of U_{TE}≥0.6 based on a detailed description of TE distribution in D. melanogaster [60] and mutation-selection balance predictions (see Materials and Methods for details). Thus, a genome-wide diploid deleterious mutation rate of ~1.2 per generation is a reasonable approximation that captures the consequences of point mutations, small indels and the insertion of transposable elements.

To assess how robust our results and conclusions are to the parameters of the BGS model, we obtained genome-wide landscapes for B under eight different models, with DDFE following a log-normal or a gamma distribution (models M_{LN} and M_{G} respectively), with deleterious mutations rates that include or not TE insertions (models M_{StdMut} and M_{LowMut} respectively), and with recombination taking into account crossover and gene conversion events or only crossovers (models MCO,GC and MCO, respectively). Unless specifically noted, we report results based on the BGS model that is most consistent with our current knowledge of gene distribution across the genome, a log-normal DDFE, a genome-wide diploid deleterious mutation rate of U=1.2, and recombination rates that include crossover and gene conversion events (i.e., our default model is M_{LN,StdMut,CO+GC}). Table S1 summarizes the results from the BGS models and Table S2 provides the full distribution of B estimates across all chromosomes.

Patterns of BGS across the D. melanogaster genome

Genome-wide estimates of B show a median of 0.591 and indicate that the predicted influence of BGS across the D. melanogaster genome would reduce the overall N_e substantially relative to levels predicted by evolutionary models with free recombination (see Figure 1A). The study of the distribution of B along chromosomes shows that the reduction in neutral diversity is severe in a large fraction of the genome, with 19% of all 1-kb regions with B<0.25 and a lower 90% CI for B of 0.605 (Figure 1B and Figure 2). Importantly, the distribution of B across trimmed chromosomes is also highly heterogenous. As expected, estimates of B are strongly influenced by variation in local crossover rates (c), with a Spearman’s rank correlation coefficients (ρ) between B and c of 0.792 for trimmed chromosomes. As shown in Figure 3, however, there is detectable variance in B for a given local c that exposes the additional effects of long-range distribution of recombination rates and genes when estimating B at a focal point.

Median B across trimmed chromosome arms is 0.643, with a minimum estimate of 0.19. Significant and variable BGS effects are, therefore, expected in D. melanogaster not only due to subtelomeric and -centromeric regions but also across trimmed chromosomes (see also [33]). This general conclusion does not vary qualitatively when considering BGS models with other parameters (Table S1). As expected, a model with a DDFE following a gamma distribution (M_{G}) predicts stronger BGS effects and lower estimates of B across the genome than when the DDFE follows a log-normal (as in our default model). Under model M_{CO,GC}, the median B is 0.428 and the lower 90% CI for B is 0.001 (median B across trimmed chromosomes is 0.493, with a minimum estimate of 0.007). Also anticipated, models with a lower deleterious mutation rate (models M_{LowMut}) generate higher estimates of B than when TE insertions are taken into account (models M_{StdMut}). For instance, median B increases from 0.591(M_{LN,StdMut,CO+GC}) to 0.769 (M_{LN,LowMut,CO+GC}), and from 0.428 (M_{G,StdMut,CO+GC}) to 0.654 (M_{G,LowMut,CO+GC}). In addition, the comparison of predictions under models with and without gene conversion shows that the standard approach of considering crossover as the only source of recombination between sites would overestimate linkage effects. Median estimates of B are 20 and 21% lower for models M_{LN,CO} and M_{G,CO} than for M_{LN,CO+GC} and M_{G,CO+GC}, respectively. The use of only crossover rates in BGS models skews estimates of B particularly in regions with intermediate rates (~0.2–2 cm/Mb), mostly across trimmed chromosomes. Both crossover and gene conversion data, therefore, need to be considered to obtain accurate estimates of the consequences of selection on linked sites and, in this case, the magnitude of BGS effects.

Finally, it is worth noting that although the different BGS models predict different point estimates and ranges of B across the chromosomes, all models generate B estimates across the genome that have virtually the same relative ranking (i.e., monotonically related; Table S3). Pairwise Spearman’s ρ between estimates of B from different BGS models range between ρ=0.9836 (comparing the two most distinct models M_{LN,StdMut,CO} and M_{G,LowMut,CO+GC}) and ρ>0.9999 (for the four comparisons between models differing in the deleterious mutation rate).

No BGS-free regions in the D. melanogaster genome

The upper end of the distribution of B across the genome is also informative and relevant for population genetic analyses of selection and demography that benefit from using regions evolving not only under neutrality but also free of linkage
under model MLN,StdMut,CO

distribution of

Box and whiskers indicate 50% and 90% CI, respectively. (B) Frequency

The median is identified by the line inside the box, the length of the

complete genome, and autosomes and the X chromosome separately.

doi:10.1371/journal.pgen.1004434.g001

Figure 1. Genome-wide estimates of BGS.

(A) Boxplots of
estimates of $B$ for complete or trimmed chromosomes and for models $M_{\text{NL,LowMut,CO-GC}} (M_{\text{CO-GC}})$ and $M_{\text{NL,LowMut,CO}} (M_{\text{CO}})$. Results shown for the complete genome, and autosomes and the X chromosome separately. The median is identified by the line inside the box, the length of the box and whiskers indicate 50% and 90% CI, respectively. (B) Frequency distribution of $B$ estimates for complete or trimmed chromosomes under model $M_{\text{NL,LowMut,CO-GC}}$. All results based on the analysis of 1-kb non-overlapping regions.

doi:10.1371/journal.pgen.1004434.g001

effects. The upper 90% CI for $B$ is 0.80 across the whole genome and 0.814 across the trimmed genome (Figures 1 and 2 and Table S1). Out of 119,027 1-kb regions investigated across the genome, the highest $B$ is 0.897 and is located in a genomic region with very low density of genes and high crossover rate; position 5,027 Mb of the X chromosome, in the middle of a large 50-kb region with a single and short CDS and a crossover rate of $\sim 14$ cM/Mb. The 1-kb autosomal region with the least BGS effects shows $B = 0.822$. BGS, therefore, plays a significant role across the entire genome, including long introns and intergenic sequences distant from genes in the middle of regions with high recombination rates. This conclusion is unlikely to be influenced by the parameters of the BGS model. As indicated, $M_G$ models predict stronger BGS effects and lower $B$ and, accordingly, we observe an upper 90% CI for $B$ of 0.707 and a maximum estimate of 0.87. On the other hand, models with lower deleterious mutation rates generate higher $B$, but even these models predict detectable BGS across the whole genome. $M_{\text{LowMut}}$ models generate upper 90% CIs for $B$ ranging between 0.811 (MLN,LowMut,CO-GC) and 0.895 (MNL,LowMut,CO-GC), and the highest 1-kb estimate of $B$ obtained by any of our $M_{\text{LowMut}}$ models is 0.947 (see Table S1 and Figure S1). That is, these results strongly suggest that while there may be a fraction of sites free of selective constraints across the $D. melanogaster$ genome, all sites might be, nonetheless, under the influence of selection at linked sites and BGS in particular.

Autosomes show stronger BGS effects than the X chromosome. The X chromosome in $D. melanogaster$ recombines at a higher rate than autosomes, caused by a higher median crossover rate per female meiosis (2.48 $\text{cM}/\text{Mb}$ for X and autosomes, respectively), and the fact that the X chromosome spends less time than autosomes in males (that do not recombine during meiosis). This higher recombination, in turn, forecasts weaker BGS in the X [30,61]. In agreement, Charlesworth [33] predicted weaker BGS effects [higher $B$] in the middle of the X chromosome than in the middle of an autosome under a number of models. Our study shows the same trend (see Figure 1A and Table S1), with median $B$ of 0.559 and 0.736 for autosomes and and the X chromosome, respectively (0.619 and 0.761 for trimmed autosomes and the X chromosome, respectively). The direct comparison of all 1-kb regions reveals that this higher $B$ in the X chromosome relative to autosomes is highly significant (Mann-Whitney $U$ Test, $P < 1 \times 10^{-12}$ for complete and trimmed chromosomes), with all BGS models generating the same trend and level of significance.

The ratio of neutral diversity for the X and autosomes (X/A) predicted by our default BGS model is 0.99 and 0.92 for complete and trimmed chromosomes, respectively. Therefore, differences in BGS effects at the X and autosomes can, at least in part, explain the observation that the X/A ratio of neutral diversity in several $D. melanogaster$ populations is higher than the 0.75 predicted by most neutral models and a 1:1 sex ratio (see [33]). Note that X/A ratios would be overestimated if gene conversion events were not taken into account, with X/A ratios of 1.12 and 0.99 for complete and trimmed chromosomes, respectively.

The genomic units of BGS in $D. melanogaster$

In this study, all sites along a chromosome were allowed to potentially play a role adding up BGS effects at any focal region of the same chromosome. To investigate the size of the genomic region causing detectable BGS effects in $D. melanogaster$, we estimated the size of the region surrounding a focal 1-kb needed to generate 90% of the total BGS effect obtained when considering the complete chromosome ($D_{B90}$ in either genetic or physical units). Equivalently, we also obtained $D_{B75}$ and $D_{B50}$ as the size of the genomic region needed to generate 75 and 50%, respectively, of the total BGS effect obtained when considering the whole chromosome.

The study of complete chromosomes shows a median genetic $D_{B90}$, $D_{B75}$ and $D_{B50}$ of 5.5, 1.2 and 0.15 cM, respectively. In terms of physical distance, the median $D_{B90}$, $D_{B75}$ and $D_{B50}$ is 2024, 477 and 76 kb, respectively (Figure 4A). Although the overall effects of BGS are reduced along trimmed chromosomes compared to whole chromosomes, the size of the region playing a significant role in the final magnitude of $B$ at a focal point is fairly equivalent, with 6.9, 1.8 and 0.21 cM for $D_{B90}$, $D_{B75}$ and $D_{B50}$, respectively (2,412, 640 and 84 kb for $D_{B90}$, $D_{B75}$ and $D_{B50}$).
This genetic and physical scale, moreover, increases with crossover rates (Figure 4B). This analysis, therefore, suggests that the extent of BGS at most genes and intergenic sites across the *D. melanogaster* genome is influenced by the cumulative effects of many sites and include numerous other genes. Thus, accurate estimates of $B$ in *D. melanogaster* require the study of genomic regions at the cM or Mb scale, ideally full chromosomes. Otherwise, BGS can be severely underestimated, and inferences about demographic events or other types of selection may be unwarranted. These results are also in agreement with the previous observation that all intergenic sequences and introns across the genome are predicted to be influenced by BGS.

Estimates of $B$ are a very strong predictor of nucleotide diversity across the whole *D. melanogaster* genome

A second goal of this study was to estimate how much of the observed levels of neutral diversity across the *D. melanogaster* genome can be explained by a BGS landscape obtained independently from variation data. A strong positive correlation would not only indicate that BGS should not be ignored in population genetic analyses but also that our estimates of $B$ are likely suitable as baseline to infer additional types of selection and/or demographic events. Because our best experimentally-obtained whole-genome recombination maps for crossover and gene

**Figure 2. High-resolution distribution of BGS effects across the *D. melanogaster* genome.** Estimates of BGS effects are measured as $B$ and shown along each chromosome arm for 100-kb adjacent windows. Red and blue lines depict estimates of $B$ based on models $M_{\text{NL,StdMut,CO+GC}}$ ($M_{\text{CO+GC}}$) and $M_{\text{NL,StdMut,CO}}$ ($M_{\text{CO}}$), respectively. Grey dashed lines show the distribution of crossover rates ($c$), measured as centimorgans (cM) per megabase (Mb) per female meiosis (see [31] for details).

doi:10.1371/journal.pgen.1004434.g002

**Figure 3. Relationship between local recombination rates and estimates of $B$ across trimmed chromosomes.** Local recombination rates ($c$) measured as cM/Mb per female meiosis. Estimates of $B$ based on the default model $M_{\text{NL,StdMut,CO+GC}}$. Results shown for 10-kb non-overlapping regions.

doi:10.1371/journal.pgen.1004434.g003
100-kb regions, \(P<1 \times 10^{-12}\), and increases up to \(\rho = 0.836\) \((P<1 \times 10^{-12})\) along individual autosome arms. Equivalent results are obtained when silent diversity is analyzed separately at intergenic and intronic sites, with \(\rho = 0.736\) between \(B\) and \(\pi_{\text{intergenic}}\) and \(\rho = 0.741\) between \(B\) and \(\pi_{\text{intron}}\) \((P<1 \times 10^{-12}\) in both cases). The study of individual autosome arms shows a positive association up to \(\rho = 0.799\) and \(0.900\) for intergenic and intronic sites, respectively \((P<1 \times 10^{-12}\) in both cases).

The predictive nature of the \(B\) landscape in \(D. \text{melanogaster}\) remains remarkably high along trimmed autosomes where BGS has been often assumed to play a minor role explaining variation in levels of polymorphism. The correlation between \(B\) and \(\pi_d\) is \(\rho = 0.529\), ranging up to \(\rho = 0.655\) when trimmed chromosome arms are analyzed separately \((P<1 \times 10^{-12}\) in all cases). Additionally, the BGS models investigated generate a stronger association between \(B\) and \(\pi_d\) than between estimates of local crossover \((\rho\) and \(\pi_{ab}\) particularly along trimmed chromosomes \((\rho = 0.677\) and \(\rho = 0.397\) for complete and trimmed autosomes, respectively). This last result exposes the limitations of using local \(c\) as an estimate of the overall strength of linked selection at a given genomic position, and highlights the importance of including long-range information of recombination rates and gene structures. Altogether, these results show the high predictive value of simple BGS models, with almost 60% of the observed variance in \(\pi_d\) across 100-kb autosomal regions explained by BGS, a percentage that is as high as \(\sim 70\%\) when investigating variation in nucleotide diversity along individual chromosome arms (see Table 1).

**Robustness to different BGS parameters.** The results presented above suggest that BGS could explain a very large fraction of the observed variation in diversity across the genome under the model that fits better our current selection and mutation data in \(D. \text{melanogaster}\) (model \(M_{\text{In,StdMut}}\)). Importantly, BGS models using different DDFEs \((M_d\) instead of \(M_{\text{In}}\) and/or a lower deleterious mutation rate \((M_{\text{LeaMul}}\) instead of \(M_{\text{StdMut}}\)) generate equivalent results. Table 1 shows \(\rho\) between \(B\) and \(\pi_d\) for the different BGS models investigated: \(\rho\) between \(B\) and \(\pi_{ab}\) ranges between 0.749 and 0.773 for complete autosomes, and between 0.514 and 0.531 for trimmed autosomes \((P<1 \times 10^{-12}\) in all cases). The study of intergenic and intronic sites separately generates similar outcomes, with \(\rho\) between \(B\) and \(\pi_{\text{intergenic}}\) ranging between 0.736 and 0.739, and \(\rho\) between \(B\) and \(\pi_{\text{intron}}\) ranging between 0.741 and 0.744 for the different models \((P<1 \times 10^{-12}\) in all cases). The similarity of outcomes should not be surprising based on the very high pairwise rank correlations between estimates of \(B\) generated by the different BGS models described above (Table S3). Taken together, these results emphasize the robustness of the approach to generate genome-wide baselines of \(B\) to study variation in nucleotide diversity along chromosomes.

**The X chromosome.** Langley et al. (2012) [26] recently showed that the correlation between crossover rates and levels of polymorphism is weaker along the X chromosome than in autosomes. In agreement, we also observe that the association between \(B\) and \(\pi_d\) along the X chromosome is weaker than in autosomes, although it is still highly significant. Estimates of \(\rho\) between \(B\) and \(\pi_{ab}\) range between 0.564 and 0.568 for the different BGS models \((P<1 \times 10^{-12}\) in all cases), and between 0.366 \((P=4 \times 10^{-7}\) and 0.373 \((P=2 \times 10^{-7}\)) for the trimmed X chromosome. Moreover, \(\pi_d\) shows a weaker correlation with crossover rates than with \(B\) along the X: \(\rho\) between \(\pi_d\) and \(c\) is 0.526 \((P<1 \times 10^{-12}\) and 0.322 \((P=9.1 \times 10^{-7}\)) for the complete and trimmed X chromosome, respectively. These results are consistent with BGS playing a weaker role along the X chromosome due to higher recombination rates [see above and [26,30,33,61]]. As discussed below, however, additional causes
might be influencing X-linked extant variation, and include higher effectiveness of selection relative to autosomes.

**Candidate regions for recent selective sweeps or balancing selection using BGS predictions as baseline**

The robustness and high predictive power of the BGS models to explain qualitative trends of nucleotide diversity across the genome, suggest that we can investigate the presence of other forms of selection by searching for regions that depart from BGS expectations. We, therefore, compared observed \( \pi_{sil} \) and levels of diversity predicted by \( B \), and parameterized departures by using studentized residuals (\( \hat{\pi}_{sil} \); see Material and Methods). Overall, the distribution of \( \hat{\pi}_{sil} \) does not show a significant departure from normality (\( \chi^2 = 28.9, \text{d.f.} = 23, P = 0.183 \)) thus validating the approach. Nevertheless, there are 58 outlier regions with nominal

**Table 1.** Correlation coefficients between estimates of \( B \) and levels of polymorphism at noncoding sites (\( \pi_{sil} \)) for different BGS models.

<table>
<thead>
<tr>
<th>BGS Model</th>
<th>Distribution</th>
<th>DDFF</th>
<th>U</th>
<th>Recombination</th>
<th>Complete Chromosomes</th>
<th>Trimmed Chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_{NL,StdMut,CO} )</td>
<td>Log-normal</td>
<td>1.2</td>
<td>CO+GC</td>
<td>0.770 (0.836)</td>
<td>0.529 (0.655)</td>
<td></td>
</tr>
<tr>
<td>( M_{NL,StdMut,CO} )</td>
<td>Gamma</td>
<td>1.2</td>
<td>CO+GC</td>
<td>0.722 (0.638)</td>
<td>0.351 (0.569)</td>
<td></td>
</tr>
<tr>
<td>( M_{NL,LowMut,CO} )</td>
<td>Log-normal</td>
<td>0.6</td>
<td>CO+GC</td>
<td>0.770 (0.836)</td>
<td>0.529 (0.655)</td>
<td></td>
</tr>
<tr>
<td>( M_{NL,LowMut,CO} )</td>
<td>Gamma</td>
<td>0.6</td>
<td>CO+GC</td>
<td>0.773 (0.838)</td>
<td>0.531 (0.659)</td>
<td></td>
</tr>
<tr>
<td>( M_{NL,StdMut,CO} )</td>
<td>Log-normal</td>
<td>1.2</td>
<td>CO</td>
<td>0.749 (0.808)</td>
<td>0.497 (0.598)</td>
<td></td>
</tr>
<tr>
<td>( M_{NL,StdMut,CO} )</td>
<td>Gamma</td>
<td>1.2</td>
<td>CO</td>
<td>0.761 (0.824)</td>
<td>0.514 (0.630)</td>
<td></td>
</tr>
<tr>
<td>( M_{NL,LowMut,CO} )</td>
<td>Log-normal</td>
<td>0.6</td>
<td>CO</td>
<td>0.749 (0.808)</td>
<td>0.497 (0.596)</td>
<td></td>
</tr>
<tr>
<td>( M_{NL,LowMut,CO} )</td>
<td>Gamma</td>
<td>0.6</td>
<td>CO</td>
<td>0.761 (0.824)</td>
<td>0.514 (0.630)</td>
<td></td>
</tr>
</tbody>
</table>

Spearman’s rank correlation coefficient (\( \rho \)) between estimates of \( B \) and \( \pi_{sil} \) (\( P < 1 \times 10^{-12} \) for all cases). Also shown, \( \rho \) between local crossover rates (\( c; \text{cm/Mb} \)) and \( \pi_{sil} \) (\( P < 1 \times 10^{-12} \), \( q \)-value <0.10), suggests a recent selective sweep at position 8.0–8.1 Mb along chromosome arm 2R. This genomic region includes gene \( Cyp9g1 \) and also showed the strongest signal of directional selection and selective sweep in large-scale population genetic analyses of North American [26] and Australian \( D. melano gaster \) [64] populations. Not all regions with a severe reduction in \( \pi_{sil} \) across the trimmed genome, however, may need recent adaptive explanations. A number of regions with \( \pi_{sil} \) much smaller than the median (e.g., 0.002 relative to a median of 0.008; see Figure 6A) also show estimates of \( B \) of 0.25 or smaller and, thus, the observed \( \pi_{sil} \) would be close to the predicted level of neutral diversity when a BGS context is taken into account.

**Uncovering the signature of balancing selection.** Numerous regions show an excess of \( \pi_{sil} \) relative to the BGS baseline and are, therefore, candidates for containing sequences experiencing balancing selection. Importantly, many of the regions showing significantly higher \( \pi_{sil} \) would not stand out unless when compared to region-specific BGS predictions. At this scale, for instance, the second strongest deviation from BGS expectations is located at position 18.4–18.5 Mb along chromosome arm 3L (\( P = 1 \times 10^{-8} \)), with a \( \pi_{sil} \) of 0.0093 that would not be noticeably high using the standard approach of comparing it to genome-wide expectations (Figure 6A). Another candidate region under balancing selection is located in cytological location 38A, showing a level of diversity also close to the average along autosomes but almost three-fold higher than the level predicted by \( B \) (\( P = 0.002 \)). This candidate genomic region with a relative excess of \( \pi_{sil} \) is within a QTL detected for variation in life span and proposed to be maintained by balancing selection in \( D. melano gaster \) [65–67]. Certainly, the detection of many regions with a relative excess of \( \pi_{sil} \) based on BGS expectations opens the possibility that molecular signatures of balancing selection may have been masked by BGS
Analyses of BGS effects and outliers of diversity at 10-kb and 1-kb scales

Because the genome-wide recombination maps used in this study were generated to have good accuracy at the scale of 100-kb, our BGS models assumed homogeneous rates within each 100-kb region. Notably, these models predict variation in B across 100-kb regions due to the heterogeneous location of genes and exons within these regions and the differential effects of proximal and distal flanking regions. Nevertheless, detailed analyses of a few small genomic regions have revealed recombination rate variation at a smaller scale in Drosophila [32,68]. Therefore, outliers from BGS expectations at scales smaller than 100-kb can reveal the localized fingerprints of other types of selection (directional or balancing selection) but the possibility of uncharacterized heterogeneity in recombination within these regions cannot be formally ruled out. That said, the study of the relevant size of the genomic region adding up BGS effects at a given focal point (with $D_{RR} > 200$ kb; see above) suggests that very local recombination variation may play a limited role influencing B at a focal point. With these caveats and considerations in mind, we investigated the presence of outliers at the scale of 10 and 1-kb to identify candidate regions under positive or balancing selection using the approach discussed for 100-kb regions.

The strong relationship between B and the observed level of silent diversity is maintained when analyzing smaller regions (Figure 5). At 10-kb scale, B remains a very good predictor of $\pi_{dil}$ along complete autosomes ($\rho = 0.678$, 8,883 regions; Figure 6B) whereas $\rho$ is 0.551 (55,467 regions) at the finest scale of 1-kb ($P < 1 \times 10^{-12}$ in both cases). The use of BGS models with different parameters (M Ln,StdMut, M G,StdMut or M Ln,LowMut) generate virtually equivalent results, with $\rho$ between estimates of B and $\pi_{dil}$ ranging between 0.678 and 0.682 for analyses at the 10-kb scale, and with $\rho$ ranging between 0.551 and 0.554 for analyses at the 1-kb scale. As observed before, B along the X chromosome shows reduced association with $\pi_{dil}$ than for autosomes also at 10- and 1-kb resolution. For X-linked regions, the correlation between B and $\pi_{dil}$ is $\rho = 0.397$ (1,979 regions) and

Table 2. Correlation coefficients between estimates of B and rates of protein evolution ($\omega_K$).

<table>
<thead>
<tr>
<th>Individual genes</th>
<th>Spearman’s $\rho$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>-0.057</td>
<td>$2 \times 10^{-12}$</td>
</tr>
<tr>
<td>Autosomes</td>
<td>-0.071</td>
<td>$5.7 \times 10^{-8}$</td>
</tr>
<tr>
<td>X chromosome</td>
<td>-0.189</td>
<td>$3.4 \times 10^{-8}$</td>
</tr>
<tr>
<td>100-kb regions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>-0.187</td>
<td>$6.6 \times 10^{-9}$</td>
</tr>
<tr>
<td>Autosomes</td>
<td>-0.160</td>
<td>$6.1 \times 10^{-6}$</td>
</tr>
<tr>
<td>X chromosome</td>
<td>-0.367</td>
<td>$1.5 \times 10^{-6}$</td>
</tr>
</tbody>
</table>

Rates of protein evolution ($\omega_K$) were obtained after controlling for selection on synonymous mutations based on residual analysis (see Materials and Methods for details). Spearman’s rank correlation coefficients ($\rho$) are shown between B and $\omega_K$ for individual genes and for the average B and $\omega_K$ for all genes within 100-kb non-overlapping regions.

doi:10.1371/journal.pgen.1004434.t002
Outliers at 10-kb scale. Among the 10,812 regions under analysis, 208 depart from expectations with nominal P<0.01, 20 showing a relative deficit of \( \pi_{\text{sl}} \) and 189 with a relative excess of \( \pi_{\text{sl}} \). Nine of these regions remain significant after correction for multiple tests (FDR q-value<0.10), and all of them show an excess of \( \pi_{\text{sl}} \) (Figure 6B). Among the candidate regions for a recent selective sweep (relative deficit of \( \pi_{\text{sl}} \)), we detect 3 genomic regions with clusters of several 10-kb regions with nominal P<0.01. In agreement with the analyses at 100-kb scale, 10 consecutive 10-kb regions show significant deficit of variation, near gene Cyph61g1 (Figure 6B). A second cluster of six 10-kb regions with reduced variation within a 120-kb interval is detected in chromosome arm 2R, with a peak signal centered at gene Dil (Distal-less), a transcription factor that plays a role in larval and adult appendage development. A third region with two adjacent 10-kb regions showing a relative deficit of \( \pi_{\text{sl}} \) is located at gene CG32783 (a protein-encoding gene with no known orthologs outside the melanogaster subgroup).

On the other hand, we detect many 10-kb regions that show a relative excess of diversity and are candidate regions for balancing selection. The strongest signal associated with excess of \( \pi_{\text{sl}} \) based on 100-kb regions now shows significantly higher variation at four adjacent 10-kb regions (19.360–19.400 Mb along chromosome arm 2L). Other strong candidate regions under balancing selection detected at this fine scale (all with FDR q-value<0.10) include the genes PH4NE2 (oxidation-reduction process), dpr20 (sensory perception of chemical stimulus), Dhc16F (regulation of transcription and microtubule-based movement) and Gai. Note that gene PH4NE2 was also shown to have an excess of protein polymorphism in the sister species D. simulans [24], thus potentially revealing a rare case of trans-specific balancing selection.

To further investigate whether the outlier regions are indeed associated with recent selective sweeps and balancing selection, we estimated Tajima’s D \( \left(D_{\text{T}}; \right. \left[69\right) \) to quantify potential differences in the frequency of polymorphic variants within the population. Selective sweeps and balancing selection are predicted to generate an excess of low-frequency (negative \( D_{\text{T}} \)) and intermediate-frequency (positive \( D_{\text{T}} \)) variants, respectively. In agreement, outlier regions with negative and positive \( \pi_{\text{sl}} \) have more negative (Mann-Whitney U-Test, \( P=4.3 \times 10^{-10} \)) and more positive (\( P<1 \times 10^{-12} \)) \( D_{\text{T}} \), respectively, than regions not departing from BGS expectations. Moreover, we observe a positive association between \( \pi_{\text{sl}} \) and \( D_{\text{T}} \) across the genome \( (\rho=0.289, P<1 \times 10^{-12}) \) that is stronger than between \( \pi_{\text{sl}} \) and \( D_{1} \) \( (\rho=0.087) \). Notably, this association between \( \pi_{\text{sl}} \) and \( D_{1} \) increases across trimmed chromosomes \( (\rho=0.302 \text{ and } \rho=0.637 \text{ for trimmed autosomes and the X chromosome, respectively; } P<1 \times 10^{-12}) \). Taken together, these results reinforce the concept that estimates of \( \pi_{\text{sl}} \) obtained when using BGS predictions as baseline are a good predictor of recent selective sweeps and balancing selection, capturing departures in number of polymorphic sites as well as the expected consequences on variant frequency.

Outliers at 1-kb scale. The study of 1-kb regions reveals 1,213 of them having diversity levels that depart from BGS expectations with nominal P<0.01, 1,186 and 27 of these regions show a relative excess and deficit of \( \pi_{\text{sl}} \), respectively. Fifty-two regions (0.076%) show an FDR corrected q<0.10, all with higher \( \pi_{\text{sl}} \) than predicted by the BGS model. Twenty-five out of the 27 1-kb regions showing a reduction in \( \pi_{\text{sl}} \) with \( P<0.01 \) cluster together at position 8.017–8.103 Mb along chromosome arm 2R, in agreement with the analyses at 100- and 10-kb scales that detected outlier regions near gene Cyph61g1. This more detailed analysis suggests that the target of selection is likely located at or proximal to Cyph61g1. The other two outlier regions with deficit in \( \pi_{\text{sl}} \) are located at genes cie (regulation of transcription) and CG11266 (mRNA binding and splicing).

In terms of regions with an excess of \( \pi_{\text{sl}} \), the strongest signal of balancing selection based on 100- and 10-kb analyses (19.3–19.4 along chromosome arm 2L) is now localized with more precision close to genes CG17349 and CG17730. Other regions detected at 10-kb scale are also identified at this finest scale and include genes PH4NE2, dpr20, and Dhc16F (all with FDR q-value<0.10). Additional candidate regions under balancing selection include genes CgaI/A/Cga/2/GoB (antibacterial humoral response), Sma-5c (olfactory behavior), CG3946 (inter-male aggressive behavior), IM4 (defense response), Cyph6a16 (a P450 gene), and three genes encoding cuticular proteins (Cpr1A, Cpr62Bb and Cpr63Ez), among others. Finally, the study of 1-kb regions also shows a strong positive relationship between \( \pi_{\text{sl}} \) and \( D_{1} \) \( (\rho=0.477, P<1 \times 10^{-12}) \), in agreement with the expected consequences of selective sweeps and balancing selection on the frequency of segregating variants (see above).

Negative relationship between estimates of \( B \) and the rate of protein evolution

Another prediction of the models of selection and linkage (either HHs or BGS) is that, parallel to a reduction in intra-specific variation, there will be a reduction in efficacy of selection (i.e., the Hill-Robinson effect [10,70–76]). This general prediction has been previously confirmed in Drosophila using local low-resolution crossover rates as indirect measure for the magnitude of Hill-Robinson effects acting on a gene. These studies showed weak but highly significant associations between crossover rates and estimates of codon usage bias or rates of protein evolution [71,74,75,77–85].

To investigate whether \( B \) landscapes also capture differences in efficacy of selection, we focused on selection against amino acid substitutions along the D. melanogaster lineage, after split from the D. simulans lineage [less than 5 mya [86]]. To this end, we obtained the ratio of nonsynonymous to synonymous changes (\( \omega_{\text{a}}, \omega_{\text{d}} = \frac{d_{\text{a}}}{d_{\text{d}}} \)) for 6,677 protein encoding genes and, more informatively, the variation in \( \omega_{\text{a}} \) after controlling for selection on synonymous mutations based on residual analysis (\( \omega_{\text{R}} \); see Materials and Methods for details). When each gene is analyzed as a single data point, there is a negative association between \( B \) and \( \omega_{\text{R}} \) \( (\rho=-0.086, P=2 \times 10^{-12}) \); Table 2). Interestingly, and contrary to the results of nucleotide diversity, the X chromosome shows a tendency for a stronger effect of \( B \) on \( \omega_{\text{R}} \) than autosomes: \( \rho=-0.189 \text{ (}\rho=3.4 \times 10^{-05}) \) and \( \rho=-0.071 \text{ (}\rho=5.7 \times 10^{-05}) \) for X-linked and autosomal genes, respectively. An equivalent but more clear pattern is observed at the scale of the resolution of our recombinations maps (100 kb), where estimating the average \( \omega_{\text{a}} \) and \( \omega_{\text{R}} \) for all genes within each region also allows for reducing idiosyncrasies of different genes influencing rates of protein evolution (e.g., gene expression breadth and levels, protein length, etc.; see [84]). At this scale, variation in \( B \) is negatively associated with \( \omega_{\text{R}} \) along autosomes \( (\rho=-0.160, P=6.1 \times 10^{-06}) \) and the X chromosome \( (\rho=-0.367, P=1.3 \times 10^{-06}) \); Table 2). Again, the association between estimates of \( B \) and rates of protein evolution is robust to different BGS models and parameters. Equivalent \( \rho \) are obtained for all eight BGS models investigated, and this is observed when analyzing individual genes (\( \rho \) between \( B \) and \( \omega_{\text{R}} \) ranging between −0.0566 and −0.0874; \( P=5 \times 10^{-12} \)) and when using the average \( \omega_{\text{R}} \) for genes within 100-kb regions (\( \rho \) ranging between −0.187 and −0.193; \( P=5.9 \times 10^{-08} \)) across the whole genome.
Temporal variation in recombination landscapes and its consequences

The association between \( B \) and rates of amino acid substitution along the \( D. \) melanogaster lineage, although highly significant in terms of associated probability, is much weaker than that for levels of polymorphism at silent sites. Heterogeneity in overall evolutionary constraints among proteins is expected to add substantial variation when investigating the consequences of \( B \) on \( \omega \) relative to studies of \( B \) on \( \pi \) because \( \pi \) is only influenced by local \( \lambda \) and the mutation rate. Nevertheless, temporally variable recombination rates at a given genomic location would also reduce the association between \( B \) and \( \omega \). Indeed, the high degree of intra-specific variation in crossover genetic maps within current \( D. \) melanogaster populations [31] together with differences in genetic maps between closely related \( Drosophila \) species [87–89] support the notion that recombination landscapes vary within short evolutionary scales, at least across trimmed chromosomes. Under this scenario, extant recombination rates and the corresponding estimates of linkage effects would be poor predictors of interspecific rates of protein evolution, even between closely related species.

In this study across the \( D. \) melanogaster genome, the use of recombination rates obtained experimentally would provide only an approximation for the relevant \( B \) along the lineage leading to \( D. \) melanogaster populations [84]. These estimates of \( B \) would be an even weaker predictor of \( \omega \) along the \( D. \) simulans lineage after split from the \( D. \) melanogaster lineage. In agreement, we observe no significant relationship between \( B \) and \( \omega \) estimated along the \( D. \) simulans lineage (\( p = -0.009 \) based on the default BGS model whereas the other BGS models generate \( p \) ranging between \(-0.014 \) and \(+0.011 \). \( P > 0.25 \) in all cases). A similar result has been obtained in comparisons of local crossover rates and rates of protein evolution between two other closely related \( Drosophila \) species, \( D. \) pseudoobscura and \( D. \) persimilis [88].

In species where BGS plays a significant role, temporal fluctuations in recombination landscapes could influence a number of analyses that assume constancy in \( \lambda \), including estimates of the fraction of adaptive substitutions (\( \alpha \) [6,90–92]). Several studies have shown that the bias in estimating \( \alpha \) can rapidly reach considerable values as a consequence of demographic changes, with \( \alpha \) being overestimated when \( \lambda \) influencing polymorphism is larger than \( \lambda \) influencing divergence (\( \lambda_{B \lambda} \geq \lambda_{D \lambda} \)). We propose that temporal changes in recombination at a given genomic position would generate an equivalent scenario, with a \( B \) influencing polymorphism (\( B_{Pol} \)) that differs from long-term \( B \) influencing divergence (\( B_{Div} \)), or the corresponding terms for local \( \lambda \). Because long term \( \lambda \) can be approximated by its harmonic mean [93], temporal fluctuations of recombination landscapes (and of local \( B \) and, therefore, local \( \lambda \)) would also predict a tendency for local \( \lambda_{B \lambda} \geq \lambda_{D \lambda} \). Such scenario would allow amino acid changes to make a larger relative contribution to divergence than to polymorphism, particularly in regions where recombination has recently increased, and thus bias estimates of \( \alpha \) upward.

Precise quantitative predictions of the potential bias in \( \alpha \) would minimally depend on the rate, magnitude and physical scale of the variation in recombination landscapes along lineages. To obtain initial insight into the effects of temporal changes in recombination rates on estimates of \( \alpha \), we investigated a rather simple and conservative scenario with forward population genetic simulations. In particular, we used the program SLIM [94] to capture the consequences of temporal changes in linkage effects on estimates of \( \alpha \) when only neutral and deleterious mutations occur along an archetypal 1-Mb region for \( D. \) melanogaster that includes 100 protein coding genes (see Materials and Methods for details). Figure 7 shows the results of estimating \( \alpha \) at selected sites under fluctuating recombination rates, with cycles of moderately high recombination for 1N generations (with \( N \) indicating the diploid population size) followed by moderately low (not zero) recombination for 3N generations. Estimates of \( \alpha \) based on models that assume constant population size [34,91,95] overestimate the true \( \alpha \) particularly when extant recombination is high (\( \alpha > 0.3 \)), with an \( \alpha = 0.15 \) when the data is combined from all temporal points. As expected, models allowing for population size change [91] generate more unbiased estimates of \( \alpha \) that show, nonetheless, a tendency upward, likely due to limitations assessing older population size changes.

Discussion

Discerning the relative contribution of different types of selection to patterns of intra- and interspecific variation is not trivial, in part because essential population and demographic parameters are often not known and the potential interactions among them are still poorly characterized. Here, we obtained the baseline of diversity across the genome predicted by BGS models completely independent of the data on nucleotide variation. The results of this study suggest that there might be no euchromatic region completely free of linkage effects to deleterious mutations in \( D. \) melanogaster. Instead, there are only genomic sites (neutral or under selection) associated with diverse degrees of BGS effects and thus under highly variable local \( \lambda \) even across the recombining regions of the genome. In this regard, the heterogeneity in local \( \lambda \) may bias population genetic estimates of selection or recombination if such variation is not taken into account. The pervasive presence of BGS has also potential consequences on demographic studies because BGS is known to generate a moderate excess of
low-frequency variants [96–102]. As a result, demographic inferences based on allele-frequencies may be skewed, with consistent patterns suggestive of a recent population expansion.

The next question investigated was how much of the patterns of variation in D. melanogaster could be explained by purifying selection alone. The results are consistent with BGS playing a major role in explaining the observed heterogeneity in nucleotide diversity across the entire D. melanogaster genome. At 100-kb scale, BGS can explain ~58% (p = 0.749–0.773 for different BGS models) of the variation of πD across the genome. The study of smaller regions reduces the statistical association between B and πD, but even when analyzing 10- and 1-kb regions, B explains ~46% (p = 0.678–0.682), and ~30% (p = 0.551–0.554) of all the observed variance in πD, respectively. These percentages increase up to ~70% (p = 0.836) for 100-kb regions, ~53% (p = 0.729) for 10-kb regions, and ~36% (p = 0.599) for 1-kb regions, across individual chromosome arms.

Importantly, these conclusions are robust to differences in parameters of the BGS model (e.g., deleterious mutation rates, DDEF, or recombination) even though the precise magnitude of BGS effects does depend on these parameters. Median estimates of B range between 0.337 and 0.769 for different BGS models, but the distribution of B along the chromosomes maintains the same rank order (pairwise B generated by the BGS models is ≥ 0.9856) and is similarly associated with the observed distribution of nucleotide diversity.

Taken together, the results and conclusions of this study imply that null expectations based on models that ignore linkage effects to deleterious mutations (or assume homogeneous distribution of Ne across the genome) are likely inaccurate in D. melanogaster, even across trimmed chromosomes. The fact that deleterious mutations are more frequent than mutations involved in balancing selection or adaptive events, together with the robustness of our conclusions to reasonable ranges of selection and mutation parameters, suggest that BGS predictions may be adequate as a baseline of diversity levels and can be used detect outlier regions subject to other selective regimes.

The use of B as baseline reveals several regions with signal of recent directional selection and associated selective sweeps, where levels of variation are lower than those predicted by BGS alone (e.g., near gene Cyp4g1 [26,64]). In agreement with expectations after a selective sweep, these regions also show an excess of variants at low frequency. A number of other regions with reduced levels of variation, however, are located in recombining genomic regions where BGS is predicted to have strong effects. These results, therefore, are consistent with a detectable but limited incidence of recent classic ‘hard sweeps’ (where diversity is fully removed near the selected sites) caused by beneficial mutations with large effects. Still it must be acknowledged that older selective sweeps, sweeps caused by weakly beneficial mutations, or ‘soft sweeps’ associated with standing genetic variation or polygenic adaptation [103,104] would not be detected in this study.

In fact, estimates of the proportion of adaptive substitutions indicate that beneficial mutations are not rare in Drosophila [6,90–92]. Moreover, analyses of nucleotide diversity around amino acid substitutions suggest that a majority of these beneficial mutations involve small effects on fitness [25,105,106] and cause detectable but very localized reduction in adjacent diversity (at the scale of 23 bp; [105,106]). Therefore, a bulk of beneficial mutations in Drosophila may be difficult to detect in genome-scans of variation but could contribute significantly to differences between species and overall adaptive rates of evolution.

BGS baselines, however, are particularly adequate to detect genomic regions under balancing selection because the predicted genomic signature of an excess of diversity may not be always evident when compared to genome-wide averages or purely neutral expectations (see Figure 6). Indeed, the use of a B baseline across the whole D. melanogaster genome provides the adequate local Ne context caused by purifying selection and corresponding linkage effects. In all, our study uncovers numerous candidate regions for balancing selection, identifying genes involved in sensory perception of chemical stimuli, antibacterial humoral response, olfactory behavior, inter-male aggressive behavior, defense response, etc. These genomic regions not only show a relative excess of polymorphic sites but also have segregating variants at higher frequency than the rest of the genome (another telltale sign of balancing selection). The results based on the study of a single population are appealing since heterogeneous environments or temporal changes predict the maintenance of local polymorphism through balancing selection [107–109], and are consistent with analyses of clines in D. melanogaster that detected the signature of local adaptation and spatially varying selection between populations [64,110,111].

Additionally, the results evidence a clear difference between autosomes and the X chromosome in terms of the consequences of variable B on levels of variation. While the X chromosome exhibits a reduced association between B and neutral diversity than autosomes, it shows a better fit between B and rates of protein evolution and efficacy of selection. These patterns are predicted by higher rates of recombination and a higher fraction of deleterious mutations with minimal role generating BGS effects in the X (see Material and Methods). Another factor possibly influencing this difference between X and autosomes is stronger efficacy of selection in the X chromosome [112]. Indeed, events of adaptive and/or stabilizing selection would distort levels of neutral diversity at linked sites beyond BGS predictions and stronger selection in the X would explain a reduced association between B and levels of diversity along the X relative to autosomes. Such combined scenario of reduced BGS effects and stronger selection would also explain a number of patterns observed in Drosophila, including an increased degree of synonymous codon usage bias [112–115], stronger purifying and positive selection acting at the level of protein evolution in X-linked genes [115,116], and the ‘faster-X’ effect [117,118] reported at both protein and gene expression levels [119–123]. Of interest will be the study of species with higher average recombination rate in autosomes than in the X, thus partially uncoupling differences in linkage effects and X-specific patterns.

Moreover, the results of this and previous studies [26,87–89] suggest that recombination rates change frequently in the Drosophila genus, and often involve differences in the distribution of recombination rates across the genome (i.e., a change of the recombination landscape). Very recent changes in recombination landscapes would uncouple extant recombination rates and polymorphism patterns generated during the last few ~Ne generations and, therefore, the reported contribution of BGS to patterns of diversity across the genome may be underestimated. On the other hand, changes in the recombination environment in species where BGS plays a significant role would also predict temporal variation in local B and impact a number of studies of selection that assume constancy of Ne along lineages or that changes in Ne should be equivalent across the genome.

Temporally variable recombination landscapes can generate, for instance, spurious evidence for multimodality in the distribution of fitness effects, or lineage-specific physical clustering of amino acid changes (such clustering has been observed among Drosophila species [124,125]). Another consequence of temporally variable recombination landscapes (and local B and Ne) would be
gene- or region-specific inequality of short- and long-term $N_e$. Regions that have recently increased in recombination would show patterns of variation suggesting population expansion or bias estimates of $z$ upward, making it less negative or even positive with no adaptive mutations [126]. Moreover, these changes in recombination landscape would also forecast substantial between-gene variation in $z$ without requiring adaptive evolution. At this point, therefore, there is the open possibility that positive estimates of $z$ in Drosophila and other species with large population size (see [6,22] and references therein) may be influenced, to an unknown degree, by temporal changes in recombination rates and landscapes. Future analyses designed to estimate population size changes or the strength and frequency of adaptive events would, therefore, benefit from including variable BGS effects across genomes and along lineages, ideally discerning local variation in BGS (and local $N_e$) from genome-wide patterns that may represent true demographic events. Genome-wide analyses may also need to consider the non-negligible presence of regions under balancing selection to prevent overestimating the extent of recent sweeps based on a relative reduction in levels of diversity.

Finally, a number of limitations and areas for future improvement need to be mentioned. In terms of selection, we investigated BGS models based on either a log-normal or a gamma DDFE [39–43,46,127]. Trying to include all possible mutations with deleterious effects across a genome into a single distribution is, however, a clear oversimplification. A combination of different DDFEs for different groups of genes and/or sites would be a more fitting approach, ideally based on a priori information independent of levels of variation (e.g., based on amino acid composition, expression levels and patterns, connectivity, etc. [84]).

To gain initial insight into the potential consequences of more realistic models, we obtained estimates of $B$ under BGS models that allow for two DDFEs (one for nonsynonymous changes and one for changes at constrained noncoding sites; see Materials and Methods for details). We then investigated whether such models would alter our conclusions, mostly in terms of the proposed adequacy of using BGS predictions as baseline to detect outliers. The results show that models with two DDFEs depart only marginally from the models with a single DDFE: rank correlations between estimates of $B$ predicted by these hybrid models and all eight previous models show $r$ ranging between 0.946 and 0.998. The association between predicted $B$ and the variation of $\pi_{sd}$ across the genome is also very high (albeit slightly lower than for models based on a single DDFE), with $r'=0.711$ at 100-kb scale and $r'=0.504$ at 1-kb scale. Notably, the comparison of outliers generated by these 2-DDFE models with those obtained by the models described above reveals few differences. For instance, 50 out of the 52 significant (FDR $q$-value<$0.10$) 1-kb outlier regions based on the default model are also predicted to be equally significant outliers by a model assuming a log-normal DDFE for nonsynonymous mutations and a gamma for deleterious noncoding mutations (the other two regions show departure with $P<0.0001$). Overall, 87.1% of the 1,213 1-kb regions showing departure at $P<0.01$ using the default BGS model are also detected as outliers ($P<0.01$) under this hybrid model. These results further support the robustness of the proposed approach to detect outlier regions based on BGS baselines.

Another line of future improvement is the physical resolution of our recombination maps and, perhaps more important, how well these maps represent the recent history of a population or species. To generate $B$ landscapes, we used recombination maps that were experimentally obtained (independent of nucleotide variation data) and capture some intra-specific variation in recombination landscapes (see [31] for details). The fact that these $B$ landscapes explain a very large fraction of the variance in nucleotide diversity across the genome suggests that these recombination maps represent quite accurately the recombination landscape in the recent history of D. melanogaster. Future recombination maps obtained from additional natural strains and populations, or under different biotic/abiotic conditions, can only improve the confidence in outlier regions and, ultimately, our understanding of selective and demographic events in this species.

Materials and Methods

General BGS expectations

The consequences of BGS at a given nucleotide position in the genome (focal point) can be described by $B$ [12–15], the predicted level of neutral nucleotide diversity under circumstances where selection and linkage are allowed ($\pi$) relative to the level of diversity under complete neutrality and free recombination ($\pi_0$). Following [15,16], we have

$$B = \frac{\pi}{\pi_0} = \exp \left( - \sum_{i=0}^{n} \frac{u_i s_i}{(s_i + (1 - s_i) r_i)^{1/2}} \right)$$

where $u_i$ is the deleterious mutation rate at the $i$-th selected site out of the $n$ possibly linked sites, $s_i$ indicates the selection coefficient against a homozygous mutation and $r_i$ is the recombination frequency between the focal neutral site and the selected $i$-th site. Note that under a BGS scenario $B$ can only be equal (no BGS effects) or lower than 1, and $B<1$ indicates strong BGS effects.

Molecular evolutionary analyses indicate variable fitness effects of deleterious mutations [44,121,127,128] and, therefore, a probability distribution of deleterious fitness effects (DDFE) of mutations at site $i$, $\phi(s_i)$ needs to be included, generating

$$B = \exp \left[ - \sum_{i=0}^{n} u_i \int_{0}^{\infty} \frac{\phi(s_i) s_i}{(s_i + (1 - s_i) r_i)^{1/2}} ds_i \right]$$

$B$, therefore, is predicted to vary across the genome as consequences of the known difference in recombination rates along chromosomes as well as due to the heterogeneous distribution of sites under selection across genomes (genes, exons, etc.). However, and because we are allowing the selection coefficient $s$ to vary according to a distribution $\phi(s_i)$, some sites under selection may have selection coefficients too small to play any BGS effect, thus not all sites under section need to be included in the study [33]. We truncate the distribution of selection coefficients at $s_T$ ($s_T=1/N_e$) because mutations with $s<s_T$ are effectively neutral and do not contribute to BGS. Different DDFEs, therefore, will generate a different fraction of deleterious mutations with $s>s_T$ and, therefore, different BGS effects (see below).

High-resolution estimates of $B$ across the whole D. melanogaster genome

We investigated a model that assumes the possibility of strongly deleterious mutations at nonsynonymous sites as well as at a fraction of noncoding sites, either untranslated genic (introns and 5’- and 3’-flanking UTRs) or intergenic regions. $B$, therefore, can be estimated at any focal neutral site of the genome as the cumulative effects of deleterious mutations at any other non-synonymous, untranslated genic and intergenic site along the same chromosome. For simplicity and based on analyses of rates of evolution between Drosophila species [22,37,38,44,91], we assumed
that the distribution and magnitude of selection on intergenic or UTR noncoding strongly selected sites \( \phi_{w_s, j}(0) \) is equivalent to that on nonsynonymous sites \( \phi_{\text{NA}}(0) \).

We can estimate \( B \) at any focal neutral site of the genome using the equation described above, \( \phi_{\text{NA}}(0) \), the actual genome annotation \( D. melanogaster \) annotation release 5.47, http://flybase.org/\) and high-definition recombination maps for crossing over and gene conversion events \[31\]. The analysis of every nucleotide sites along a single chromosomal arm taking into account all other sites under selection of this same chromosome would, however, require \( >1 \times 10^{13} \) pair-wise nucleotide comparisons, each one requiring a numerical integration. To speed up the process we followed \[17\] and first obtained the integral

\[
\int_{x_j}^{1} \frac{\phi_{\text{NA}}(s_{\text{NA}}) s_{\text{NA}}}{(s_{\text{NA}} + (1 - s_{\text{NA}}) r_j)^2} ds_{\text{NA}}
\]

along a continuum for possible recombination rates between two sites, in our case for \( r \) ranging between \( 1 \times 10^{-10} \) (equivalent to \( \varepsilon = 0.01 \)) to \( 1 \); we assumed \( r = 0 \) whenever the recombination distance between two nucleotides is smaller than \( 1 \times 10^{-10} \). Because he recombination distance between any pair of nucleotides can be obtained, the generation of complete maps of \( B \) is now more tractable although still computationally very intensive. We then made the simplifying assumption of ignoring variation within 1-kb windows when estimating \( B \) at any another 1-kb window along the chromosome. That is, for the purpose of generating BGS effects, all sites within a 1-kb window have an equivalent recombination rate with sites within any other 1-kb window along the chromosome. This is a reasonable approximation at this time due to the fact that the resolution of our best genome-wide recombination maps is 100-kb and differences in recombination due to a few nucleotides play a minor role when comparing sites separated by tens of hundreds of kb.

By dividing a chromosome arm into \( L \) adjacent 1-kb windows, \( B \) at a focal neutral site located at the center of the \( j \)-th window \( (B_j) \), can be estimated by using:

\[
X_j = \sum_{i=1}^{L} \text{Nat}_i \text{Nat}_j \int \frac{\phi(s_{\text{NA}}) s_{\text{NA}}}{(s_{\text{NA}} + (1 - s_{\text{NA}}) r_j)^2} ds_{\text{NA}}
\]

\[
Y_j = \sum_{i=1}^{L} \text{Nutr} \text{Nat}_j \text{Nutr} - s_{\text{Nat}} \int \frac{\phi(s_{\text{NA}}) s_{\text{NA}}}{(s_{\text{NA}} + (1 - s_{\text{NA}}) r_j)^2} ds_{\text{NA}}
\]

\[
Z_j = \sum_{i=1}^{L} \text{Nuc} \text{Nat}_j \text{Nuc} - s_{\text{Nat}} \int \frac{\phi(s_{\text{NA}}) s_{\text{NA}}}{(s_{\text{NA}} + (1 - s_{\text{NA}}) r_j)^2} ds_{\text{NA}}
\]

\[
B_j = \exp - (X_j + Y_j + Z_j)
\]

where \( \text{Nat} \), \( \text{Nutr} \), and \( \text{Nuc} \) are the number of nonsynonymous, UTR and intergenic sites possibly under strong selection within window \( i \), respectively, and \( r_i \) is the recombination between the center of the focal window \( j \) and the center of window \( i \), with \( r_j \) increasing in 1-kb intervals. \( s_{\text{NA}} \), \( s_{\text{Nutr}} - s_{\text{Nat}} \), and \( s_{\text{Nuc}} - s_{\text{Nat}} \) are the deleterious mutation rate at nonsynonymous, UTR and intergenic sites, respectively. Because all sites in the genome are actually being taken into account, this approach avoids the need to interpolate \( B \) estimates and generates a very detailed \( B \) landscape.

**Deleterious mutation rates.** Initial estimates of the mutation rate based on mutation accumulation lines in \( D. melanogaster \) suggested a mutation rate \( \mu \) of \( \sim 8 \times 10^{-9} \) per bp/generation, for point mutations and small indels \[47\]. More recent analyses suggest a \( \sim 4 \times 10^{-8} \) bp/generation for removing data from a line with unusually high mutation rates \[49-50\]. Based on the fraction of conserved sites in exons and noncoding sequences \[22,37\], this lower mutation rate predicts a diploid rate of deleterious mutations per generation \( U \) of 0.6 for euchromatic regions. \( U \sim 0.6 \), however, is an underestimate of the true \( U \) in \( D. melanogaster \) because it does not take into account the recent bottleneck experienced by North America populations \[48\] or the deleterious consequences of TE insertions within genes and regulatory sequences. Additionally, the elevated mutation rate observed in one line \( \text{line 33} \[47,48\] \) is not a new trait developed during the mutation accumulation experiment and, therefore, such genotype was present in the initial natural population of \( D. melanogaster \). We, thus, use \( U = 0.6 \) as a lower boundary for the true deleterious mutation rate across the euchromatic genome (in models \( M_{\text{LowMut}} \)).

TEs are pervasive across the \( D. melanogaster \) genome \[51-60\]. Cridland et al. \[3\] reported a detailed analysis of TE abundance and distribution in \( D. melanogaster \) based on deep coverage, whole-genome sequencing \[60\]. Their analysis of the Drosophila Synthetic Population Resource (DSPR; \[129\]) showed a total of 7,104 TE insertions, with 633 TEs inserted within exons. To include the deleterious consequences of TE insertion, therefore, we obtained an approximate mutation (insertion) rate \( \mu_{\text{TE}} \) based on mutation-selection balance predictions and the presence of TEs in genomic regions where they are most likely to cause deleterious effects \( \text{(i.e., within exons)}. \) We use TE presence in the DSPR lines instead of analyses based on the Drosophila Genetic Reference Panel (DGRP) due to the higher sequencing coverage in the DSPR lines, which increases the likelihood to detect and map TE insertions \[60\].

For non-recessive strong deleterious mutations, mutation-selection balance predicts a probability of segregation \( P_{\text{Seg}} \), that is the product of the equilibrium frequency \( q \) and the number of chromosomes analyzed \( n \). \( P_{\text{Seg}} \) in exons \( \sim \) 0.00003 (633 TEs/21,000,000 exonic bp) and \( q = P_{\text{Seg}} / n \sim 2 \times 10^{-6} \). Based on the very low frequency of segregating TEs in natural populations \( \text{(i.e., most TEs in this study were present in a single genome in agreement with previous analyses of \( D. melanogaster \) populations; see \[53,57,58\] \)}, we can assume non-recessive deleterious effects and a heterozygous selection coefficient \( \mu_{\text{TE}} \) against TE insertions within exons to be equal or greater than for amino acid mutations \( \mu_{\text{AA}} \), so \( \mu_{\text{TE}} \) can be estimated by \( \mu_{\text{TE}} = q \mu_{\text{AA}} \). We thus obtain \( U_{\text{TE}} \sim 0.6 \), a result in agreement with previous analyses of TE transposition rate that suggested that the genomic rate of TE transposition is of the same order as the nucleotide spontaneous mutation rate \[55\], and suggests \( U \sim 1.2 \) when including point mutations, small indels and TE insertions across the euchromatic regions of the \( D. melanogaster \) genome.

**Distribution of deleterious effects.** The true probability distribution of selection coefficients against newly arising mutations is not known. Sensibly, no single deleterious distribution of fitness effects (DDFE) is likely to capture mutations at different genes or sites within genes. Different spatial and temporal conditions will, likely, further alter such DDFEs. As an approximation, we studied two DDFEs and the corresponding parameters proposed for \( Drosophila \) to partially capture the potential effects of different DDFEs on our results and conclusions.
gamma distribution [39–44] and a log-normal distribution [45,46], with the latter allowing to capture the presence of lethal mutations and fitting better D. melanogaster polymorphism data [45,46].

Following Loewe and Charlesworth [45], the log-normal DDFE has a median \( s = 0.000231 \) and standard deviation 5.300 for autosomes. A gamma DDFE for Drosophila is described by a shape parameter \( k = 0.3 \) and mean \( \mu = 0.0025 \) [33,30]. In both cases, we assumed a dominance coefficient \( h \) of 0.5. For models that include the contribution of TE insertions (\( U=1.2 \)), we further assumed that the DDFE of TEs is the same than of point mutations, and that the genomic sites where a TE insertion has deleterious effects are the same where point mutations and small indels have deleterious effects. Note that under the models that include a log-normal DDFE, the median \( s \) against individual TE insertions (0.000231) is equivalent to previous estimates in D. melanogaster of 0.0002 [54].

An interesting attribute of the log-normal distribution is that it predicts a higher frequency of extreme values, including effectively 0.0002 \((0.000231)\) is equivalent to previous estimates in D. melanogaster [33]. We assume 0.75 as the fraction of coding sites that alter amino acid sequences and correct this fraction by the proportion of constrained nonsynonymous sites \( c_s \) to focus only on the fraction of deleterious mutations. In D. melanogaster, \( c_s \) for amino acid sites is \( \approx 0.92 \) [22,37] and, thus, \( N_{ac} \) for a region is \( L_{med} \approx 0.75 \times 0.92 \). To obtain \( N_{ac,S} \), we correct the number of noncoding genic sites using the proportion of constrained sites \( 0.56 \) for introns and 0.81 for flanking UTRs [22] and, equivalently, we use the proportion of constrained sites at intergenic sequences \( \approx 0.5 \) [22,37] to obtain \( N_{gi,S} \).

The deleterious mutation rate per bp \( (u) \) will be different for different models, due to the overall mutation rate and because different DDFEs predict a different fraction of deleterious mutations with \( u<\gamma_1 \) that will not contribute to BGS. The two mutation rates investigated here, \( U=0.6 \) (MLowMut) and 1.2 (MStdMut), represent a neutral mutation rate of \( 4.2 \times 10^{-9} \) and \( 8.4 \times 10^{-9} \) per bp/generation. Assuming \( N_c = 1,000,000 \) for D. melanogaster, the log-normal and gamma DDFEs described above predict 15.3 and 7.4% of deleterious mutations with \( u<\gamma_1 \), respectively. Therefore, the deleterious mutation rate relevant for BGS is 84.7 \( (M_{LN}) \) and 92.6% \( (M_G) \) of that of the mutation rate.

**BGS models with two DDFEs.** All previous BGS models assume a single DDFE for sites under selection. As a first approximation towards more realistic description of selection across the genome, we investigated two additional models that allow for two DDFEs, one for nonsynonymous mutations and the other for deleterious noncoding mutations (each DDFE following the parameters indicated above). The \( B \) landscapes generated by these hybrid models are intermediate relative to the other eight models: a model assuming a log-normal DDFE for nonsynonymous mutations and a gamma for deleterious noncoding mutations \( (M_{LN/G,StdMut,CO+GC}) \) shows a genome-wide median \( B \) of 0.469 whereas a model with a gamma DDFE for nonsynonymous mutations and a log-normal for deleterious noncoding mutations \( (M_{G/LN,StdMut,CO+GC}) \) shows a median \( B \) of 0.459. Spearman’s \( \rho \) between estimates of \( B \) across the genome from these 2-DDFE models and the other eight single-DDFE models range between 0.974 and 0.998 for \( M_{LN/G,StdMut,CO+GC} \) and between 0.946 and 0.986 for \( M_{G/LN,StdMut,CO+GC} \).

**X chromosome.** For the X chromosome, we have assumed equivalent deleterious mutations rates than for autosomes and adjusted the distribution of selection coefficients to \( \gamma_1=2/3 \ \gamma_s \) (assuming equal number of breeding males and females in the population and genic selection, \( h=0.5 \)). Estimates of the ratio of observed neutral diversity for the X and autosomes \( (X/A) \) also assume equal number of males and females, with \( X/A = 0.75 \times B_X/B_A \). Equivalently, \( s_X \) for the X chromosome varies from that for autosomes, with \( s_X \) for the X chromosome now satisfying \( N_{ac,X s_X} \). For the X chromosome, the log-normal and gamma DDFEs predict 18.6 and 9.1% of deleterious mutations with \( u<\gamma_X \), respectively (both percentages higher that for autosomes, see above). Finally, we allowed for an effective difference in recombination rates due to the lack of meiotic recombination in Drosophila males, with 0.5r and 0.6r for autosomes and the X chromosomes, respectively. For each of the BGS models under study, we obtained specific solutions to the integrals described above along a continuum of recombination rates for the X chromosome and autosomes separately.

**Complete and trimmed chromosomes.** All analyses were carried out based on the study of complete chromosome arms as
well as after removing euchromatic regions near to centromeres and telomeres with evident reduction in crossover rates [133]. Following [26] we use the term trimmed chromosomes or trimmed genome to designate genomic regions after removing sub-centromeric and sub-telomeric regions with strongly reduced crossover rates. Sub-centromeric regions with reduced crossover rates were assigned by starting at the centromere and moving into the chromosome arm until a minimum of 3 consecutive 100-kb windows showed crossover rates >1 cM/Mb. Sub-telomeric regions with evident reduction in crossover rates were assigned in an equivalent manner, by starting at the telomere and moving towards the center of the chromosome until a minimum of 3 consecutive 100-kb windows showed >1 cM/Mb.

Estimates of neutral diversity across the D. melanogaster genome

Nucleotide diversity across the D. melanogaster genome was estimated from a sub-Saharan African population (Rwanda, RG, population of the Drosophila Population Genetics Project, DPGP; www.dpgp.org/ and [62]). D. melanogaster is thought to have originated in sub-Saharan Africa, and eastern Africa—including Rwanda—in particular [134]. Our use of the RG population, therefore, minimizes the non-equilibrium effects caused by recent expansion observed in western Africa and non-African D. melanogaster populations [62,134]. Additionally, the RG population combines a relatively large sample (n = 27), and low and well characterized levels of admixture [62]. We followed Pool et al. (2012) and used updated assemblies from the DGDP2.v3 site (see http://www.dpgp.org/dpgp2/DPGP2.html and [62] for details). We analyzed these assemblies with 1) sites putatively heterozygous or with quality value smaller than Q31 masked to ‘N’, and 2) putatively admixed regions of African genomes filtered to ‘N’ based on the description of admixed regions from [62] (also available from http://www.dpgp.org/dpgp2/DPGP2.html). Finally, we only investigated non-N sites that were present in a minimum of 10 sequences. Equivalent results were obtained when the analyses were restricted to sites with a minimum of 15 sequences, or after removing regions showing excess admixture as well as regions showing excess long-range identity-by-descent (IBD) [62] (data not shown).

Neutral (silent) diversity was estimated as pairwise nucleotide variation per site at intergenic sequences and introns (π SD) and following the approach described in [31]. In short, we first annotated all gene models, transposable elements and repetitive sequences onto the reference sequence allowing for overlapping annotation. Intergenic sites correspond to sites between gene models (excluding annotated UTRs), transposable elements or repetitive sequences. Intronic sites are analyzed as such only when they never overlap with another annotation (e.g., with alternatively spliced exons or other elements within introns). After this filtering approach, intergenic and intronic sites show similar levels of silent diversity, with a very weak tendency for π SD in introns (π introns = 0.0085) to be higher than in intergenic (π intergenic = 0.0082) sites (Sign test, P = 0.034 and P = 0.056, along autosomes and X chromosome, respectively).

Unless indicated otherwise, diversity analyses were performed using all 100-kb non-overlapping windows across the genome whereas analyses of 10-kb and 1-kb regions were limited to regions with more than 1,000 and 500 silent sites, respectively. Inferference about the frequency of segregating variants within the population were based on estimates of Tajima’s D [69] after normalizing by Dmin (D/Dmin) following Schaeffer (2002) [135]. Equivalent conclusions were obtained when using DGRP sequenced strains [136] from a North American natural population (Raleigh, NC, USA).

Outliers of nucleotide diversity. To parameterize departures of observed levels of diversity π SD relative to the levels predicted by B, we applied a generalized regression model (GRM), and obtained regression residuals π SD.R. In particular, we obtained studentized residuals to detect outlying values and obtain associated probabilities. Studentized residuals are equivalent to standardized residuals (the ratio of the residual to its standard error) and have a Student’s t distribution, but in the case of studentized residuals the data point or observation under study (possibly an outlier) is removed from the analysis of the standard error. Because the sample size in our analyses is very large and the fraction of outliers small, studentized and standardized residuals are almost equivalent and generate the same outlier regions at the three physical scales analyzed. We applied GRM separately to autosomes and X chromosome to prevent including assumptions about the true X/A ratio. To correct for multiple tests, we applied the Benjamini and Hochberg (1995) method with FDR = 0.10.

Estimates of rates of protein evolution

Genes and sequence alignments. We obtained CDS sequence alignments in the 5 species of the D. melanogaster subgroup (D. melanogaster, D. simulans, D. sechellia, D. yakuba and D. erecta) based on multiple-species alignments available from the UCSC Genome Browser [http://hgdownload.cse.ucsc.edu/ goldenPath/dm3/multiz15way/alignments/flyBaseGene.exonNuc.fa.gz] and current gene annotation and genomic location in D. melanogaster [http://flybase.org; D. melanogaster annotation release 5.47, 10/9/2012]). We removed genes with coding sequences in D. melanogaster shorter than 450 bp or with fewer than 100 amino acids positions present in all five species after alignment. We also removed genes presenting premature stop codons in at least one species relative to the stop codon position in D. melanogaster. Finally, for genes with multiple transcript forms we chose the CDS alignment corresponding to the longest transcript. In total, we investigated rates of evolution in 6,876 protein-coding genes.

Rates of protein evolution. We obtained dS (the number of nonsynonymous substitutions per nonsynonymous site), dS/dD (the number of synonymous substitutions per synonymous site) and ω (the ratio dS/dD) for each gene. To this end, we used the program codeml as implemented in PAML v 4.5 [137,138] and applied a branch model that allowed different ω in all internal and external branches of the five-species tree. We focused on ω along the D. melanogaster branch after split from the D. simulans lineage to investigate recent patterns of efficacy of selection on amino acid changes and the possible association with BGS effects across the genome based on recombination rates estimated in D. melanogaster.

We followed Larracuente et al. (2008) [84] to prevent combining genes with amino acids evolving under positive selection and genes with most amino acids evolving under a nearly neutral scenario, either purifying selection or neutrality. We, therefore, applied codeml and compared Model M1a (nearly neutral evolution) against a model that allows the additional presence of positive selection at a fraction of sites (model M2a). In particular, we compared maximum likelihood estimates (MLEs) under these two models and applied a likelihood ratio tests (LRTs) [139,140] to detect differences between the models. A total of 125 genes showed evidence of positive selection in the D. melanogaster lineage after applying the Benjamini and Hochberg (1995) method to correct for multiple tests (FDR = 0.05). To further reduce the incidence of genes under positive selection and/or recent pseudogenization we removed genes showing ω greater than 0.75 (note that the median ω in the D. melanogaster lineage is
Equivalent results are obtained when we used regression residuals of an estimate of variable efficacy of selection on amino acid changes. In D. melanogaster, synonymous mutations are under weak selection ([113] and references therein) and therefore a reduction in efficacy of selection is predicted to increase \( d_h \), possibly biasing the direct use of the ratio \( d_s/d_H \). On the other hand, the inclusion of \( d_s \) corrects for possible differences in coalescent times and ancestral polymorphism across the chromosome. We then applied generalized linear models, GRM, and obtained regression residuals of \( \omega \) along the D. melanogaster lineage after controlling for \( d_h (\omega v) \) and used \( \omega v \) as an estimate of variable efficacy of selection on amino acid changes. Equivalent results are obtained when we used regression residuals of \( d_h \) along the D. melanogaster lineage after controlling for \( d_s (\omega v) \) as an estimate of variable efficacy of selection on amino acid changes. We obtained regression residuals of \( \omega \) or \( d_h \) for autosomes and the X chromosome separately.

Forward computer simulations and estimates of the proportion of adaptive substitutions

We used the program SLIM [94] to capture the consequences of temporal changes in recombination rates on estimates of \( \alpha \). Simulations followed a panmictic population of 10,000 diploid individuals \( N \) and a chromosome segment of 1 Mb that contained 100 protein encoding genes evenly distributed, one every 10 kb. Each 10-kb region included a typical Drosophila gene: a 1,000 bp 5’ UTR, a first short 300-bp exon, a 1,000 bp first intron, two additional 600 bp exons, a short 200-bp internal intron, and a 300 bp 3’-UTR, followed by a 5,000 bp intergenic sequence. Mutations were assigned to have the same parameters than those in the BGS models described above, with two mutation types: neutral and deleterious. The mutation rate per generation was set to \( \lambda_s = 0.005 \) and deleterious mutations were assumed to follow a gamma DDFE (\( \theta = 0.5 \)) with average \( N_s = -2,500 \) [33,38]. The proportion of deleterious mutations at the different genomic elements was set to 0.92 at first and second codon positions, 0.81 at UTRs, 0.56 at introns, and 0.5 at intergenic sites [22,37]; all third codon positions evolved neutrally. Note that the simulation of a smaller population would prevent the study of deleterious mutations with the appropriate scaled selection for Drosophila (\( N_s = -2,500 \)) while the simulation of shorter genomic sequences would severely underestimate BGS effects in regions with non-reduced recombination.

Simulations followed 10 independent populations a minimum of 50 \( N \) generations after reaching equilibrium (>10 \( N \) generations). Recombination rates were uniformly distributed, with a population crossover rate (\( /\text{bp} \)) of \( N_{CG} = 0.04 \) and \( N_{CO} = 0.0025 \) for periods of high and low recombination, respectively. To prevent overestimating BGS effects we also included a constant rate of gene conversion initiation (\( /\text{bp} \)) of \( N_c = 0.05 \) and average gene conversion tract length of \( L_c = 518 \) [31]. After the initial 10 \( N \) generations, we sampled the population every 0.1 \( N \) generations, obtained polymorphism data from 20 randomly drawn chromosomes, and compared them to a sequence that evolved independently for 20 \( N \) generations to obtain divergence (\( d \) values. These levels of polymorphism and divergence are equivalent to those observed within D. melanogaster for polymorphism and between D. melanogaster-D. simulans for divergence, where \( d/H = -6 \) for neutral sites [95]. Estimates of \( \alpha \) at selected sites were restricted to the 100-kb central region to capture long-range BGS effects. Following Eyre-Walker and Keightley [91], \( \alpha \) was obtained by maximum likelihood (ML) to capture the presence of strongly deleterious mutations in the simulations, with and without the possibility of population size change, and with and without correcting for the contribution of polymorphism to divergence [95]. Estimates of \( \alpha \) were obtained using the DFE-alpha server (http://lanner.cap.ed.ac.uk/~cang33/df-alpha-server.html).

Supporting Information

Figure S1 Frequency distribution of estimates of \( B \) from BGS models that differ in the distribution of deleterious fitness effects (DDFE) and deleterious mutation rate. (A) Estimates of \( B \) based on models M\( _{LN,StdMut} \) (log-normal DDFE) and M\( _{G,StdMut} \) (gamma DDFE) when the diploid deleterious mutation rate is \( U = 1.2 \). (B) Estimates of \( B \) based on models M\( _{LN,LowMut} \) (log-normal DDFE) and M\( _{G,LowMut} \) (gamma DDFE) when the diploid deleterious mutation rate is \( U = 0.6 \) (see text for details). All results based on the analysis of 1-kb non-overlapping regions.

(TIF)

Figure S2 Distribution of silent diversity (\( \pi_s \)) and predicted BGS effects (\( B \)). Estimates of \( B \) based on model M\( _{LN,StdMut,CO+GC} \) Results shown for 100-kb non-overlapping regions across chromosome arm 3L.

(TIF)

Table S1 Summary of \( B \) estimates from the different BGS models. (PDF)

Table S2 Complete distribution of \( B \) estimates from the different BGS models. Table shows estimates of \( B \) based on eight BGS models, along all chromosome arms, and for adjacent 1-kb regions.

(XLSB)

Table S3 Pairwise Spearman’s rank correlation coefficients (\( \rho \)) between estimates of \( B \) from different BGS models. (PDF)

Acknowledgments

The author thanks Ana Llopart, Andrew Adrian and Brian Charlesworth for comments and suggestions on the manuscript, John Pool for making publicly available sequences from the Rwanda RG population (DPGP2), and the editor and three reviewers for their constructive comments that greatly improved the manuscript.

Author Contributions

Conceived and designed the experiments: JMC. Performed the experiments: JMC. Analyzed the data: JMC. Contributed reagents/materials/analysis tools: JMC. Wrote the paper: JMC.

References


