Peer Community Journal

Section: Evolutionary Biology

RESEARCH ARTICLE

Published
2023-04-26

Cite as
Gustavo V Barroso and Julien Y Dutheil (2023) *The landscape of nucleotide diversity in* Drosophila melanogaster *is shaped by mutation rate variation* , Peer Community Journal, 3: e40.

Correspondence
gvbarroso@gmail.com

Peer-review
Peer reviewed and recommended by PCI Evolutionary Biology, https://doi.org/10.24072/pci.evolbiol.100636

The landscape of nucleotide diversity in *Drosophila melanogaster* is shaped by mutation rate variation

Gustavo V Barroso [1,2] and Julien Y Dutheil [1,3]

Volume **3** (2023), article e40

https://doi.org/10.24072/pcjournal.267

Abstract

What shapes the distribution of nucleotide diversity along the genome? Attempts to answer this question have sparked debate about the roles of neutral stochastic processes and natural selection in molecular evolution. However, the mechanisms of evolution do not act in isolation, and integrative models that simultaneously consider the influence of multiple factors on diversity are lacking; without them, confounding factors lurk in the estimates. Here we present a new statistical method that jointly infers the genomic landscapes of genealogies, recombination rates and mutation rates. In doing so, our model captures the effects of genetic drift, linked selection and local mutation rates on patterns of genomic variation. We then formalize a causal model of how these microevolutionary mechanisms interact, and cast it as a linear regression to estimate their individual contributions to levels of diversity along the genome. Our analyses reclaim the well-established signature of linked selection in *Drosophila melanogaster*, but we estimate that the mutation landscape is the major driver of the genome-wide distribution of diversity in this species. Furthermore, our simulation results suggest that in many evolutionary scenarios the mutation landscape will be a crucial factor shaping diversity, depending notably on the genomic window size. We argue that incorporating mutation rate variation into the null model of molecular evolution will lead to more realistic inferences in population genomics.

[1] Max Planck Institute for Evolutionary Biology. Department of Evolutionary Genetics. August-Thienemann-Straße 2 24306 Plön – Germany, [2] University of Wisconsin, Madison. Department of Integrative Biology. 447 Birge Hall – Madison, WI – USA, [3] Unité Mixte de Recherche 5554 Institut des Sciences de l'Evolution, CNRS, IRD, EPHE, Université de Montpellier, Montpellier, France

e-ISSN 2804-3871

Introduction

Understanding how various evolutionary mechanisms shape nucleotide diversity – typically measured as the average pairwise heterozygosity, π – is a major goal of population genomics (Charlesworth, 2010; Ellegren & Galtier, 2016), with a rich history of theoretical and empirical studies that have the fruit fly *Drosophila melanogaster* as its centerpiece (Casillas & Barbadilla, 2017; Charlesworth & Charlesworth, 2017; Haudry et al., 2020). For many years, the debate focused on the relative importance of genetic drift and natural selection to the genome-wide average π (Kimura, 1968; Ohta, 1992). The observation that π does not scale linearly with population size across species (Lewontin, 1974) was termed "Lewontin's Paradox", and recent work has taken a new stab at this old problem by modeling the effect of natural selection (Buffalo, 2021; Galtier & Rousselle, 2020). Later on, with recognition that linkage and recombination wrap the genome in regions of correlated evolutionary histories (Hudson, 1983; Hudson & Kaplan, 1985), focus shifted toward understanding how diversity levels vary along chromosomes of single species (Pouyet & Gilbert, 2021). In 1992, Begun and Aquadro found a positive correlation between π and local recombination rate in *D. melanogaster*, (Begun & Aquadro, 1992) which was interpreted as the signature of linked selection (Cutter & Payseur, 2013; Hudson & Kaplan, 1988) – at first in terms of selective sweeps (Smith & Haigh, 1974; Stephan et al., 1992; Wiehe & Stephan, 1993) and soon re-framed in the light of background selection (Charlesworth et al., 1993; Hudson & Kaplan, 1995, 1994; Nordborg et al., 1996). In the three decades since these seminal works, identifying the drivers of the genome-wide distribution of diversity became a leading quest in the field of population genetics. Nevertheless, this search has so far been incomplete. The literature has mostly considered how patterns of diversity are affected by selection (Andolfatto, 2007; Comeron, 2014; Elyashiv et al., 2016; McVicker et al., 2009; Murphy et al., 2022) or introgression (Hubisz et al., 2020; Stankowski et al., 2019), whereas spatial variation in *de novo* mutation rates (μ) has been largely ignored as an actual mechanism of variation in π along the genome, presumably due to challenges in its estimation (Besenbacher et al., 2019; Jónsson et al., 2018). Yet a study based on human trios advocates that the impact of the mutation landscape on polymorphism may be greater than previously recognized: up to 46% of the human-chimpanzee divergence, and up to 69% of within-human diversity, can potentially be explained by variation in *de novo* mutation rates at the 100 kb scale (Smith et al., 2018). It is unclear, however, how well these results generalize to species with distinct genomic features and life history traits. The few studies conducted in non-human organisms relied on proxies of the local mutation rate, such as synonymous diversity or divergence with a closely-related outgroup (Castellano et al., 2020, 2018). Still, these indirect measures of the mutation rate are susceptible to the confounding effect of selection, which can act both directly (e.g. codon usage (Lawrie et al., 2013; Machado et al., 2020)) and indirectly (e.g. recent background selection in the case of synonymous diversity (Charlesworth et al., 1993; Hudson & Kaplan, 1995; Nordborg et al., 1996) as well as background selection in the ancestral population in the case of synonymous divergence (Phung et al., 2016)). Therefore, developing dedicated statistical methods to infer mutation rate variation from polymorphism data is of high interest. Through simultaneous inference of the genomic landscapes of genetic drift, linked selection, recombination and mutation, confounding factors can be better teased apart and, subsequently, the relative contribution of each of these micro-evolutionary mechanisms to the distribution of diversity can be more meaningfully quantified.

Disentangling the effects of multiple factors shaping the evolution of DNA sequences is challenging because different mechanisms can produce similar phenomena (*sensu* (Baetu, 2019)). For example, a genomic region with reduced nucleotide diversity (relative to some baseline reference) can be causally explained by either linked selection, drift, low mutation rate or a combination thereof. In an elegant effort to tease these mechanisms apart, Zeng and Jackson developed a likelihood-based model that jointly infers the effective population size (N_e) (Charlesworth, 2009) and μ with high accuracy in different parts of the genome (Zeng & Jackson, 2018). However, since it relies on the single-site frequency spectrum, their method is restricted to unlinked loci. While this approach avoids the confounding effect of linkage disequilibrium in the inference procedure (Slatkin, 2008), it discards sites in the genome where local variation in the mutation rate may be relevant as well as dismisses the gradual impact of recombination and linked selection on spatial variation in diversity. In this article, we put forward a new model to fill in this gap. We have previously described a statistical framework (the integrative sequentially Markovian coalescent, iSMC) that jointly infers the demographic history of the sampled population together with variation in the recombination rate along the genome via a Markov-modulated Markov process (Barroso et al., 2019). We now extend this framework to also account for sequential changes in the mutation rate. This integration allows statistical inference of variation along the genome in both recombination and mutation rates, as well as in Times to the Most Recent Common Ancestor (τ), that is, the ancestral recombination graph of two haploids (Rosenberg & Nordborg, 2002). Whereas drift causes stochastic fluctuations in τ around its expected value under neutrality (in diploid organisms, $E[\tau] = 2 \times Ne$), natural selection disturbs τ away from its neutral distribution near functionally constrained regions of the genome (Palamara et al., 2018; Rasmussen et al., 2014; Stern et al., 2019; Zeng & Charlesworth, 2011). Thus, iSMC offers estimators of all relevant micro-evolutionary mechanisms, and we can further use causal inference (Pearl & Mackenzie, 2018) to simultaneously estimate their effects on diversity. Our analyses of *D. melanogaster* genomes reveal the impact of linked selection; however, they suggest that the rate of *de novo* mutations is quantitatively the most important factor shaping nucleotide diversity in this species.

Methods

Modeling variation in the mutation rate along the genome

We now introduce our approach to modeling the mutation landscape starting from the original pair-wise SMC process. Because iSMC models pairs of genomes, the genealogies underlying each orthologous site can be conveniently summarized by τ, the time to their most recent common ancestor (Li & Durbin, 2011; Schiffels & Wang, 2020). The pair of DNA sequences is described as a binary string where 0 represents homozygous states and 1 represents heterozygous states (thus, once haploid genomes are combined into diploids, phasing information is discarded). The probability of observing 0 or 1 at any given position of the genome depends only on τ and the population-scaled mutation rate $\theta = 4 \times Ne \times \mu$. If the hidden state configuration of the model excludes spatial variation in the mutation rate, then θ is assumed to be a global parameter such that the emission probabilities of homozygous and heterozygous states can be computed for every site as $P(0|\tau) = e^{(-\theta \times \tau)}$, and $P(1|\tau) = 1 - e^{(-\theta \times \tau)}$ respectively, as originally presented by Li & Durbin (2011).

We estimate the per-site, genome-wide average $\hat{\theta}_0$ as the average number of pair-wise differences observed between all pairs of genomes. Therefore, the effective population size implicit in $\hat{\theta}_0 = 4 \times Ne \times \mu$ is the average of N_e along the genome, accounting for selective effects. We fix $\hat{\theta}_0$ to this point estimate and exclude it from the optimization step conducted with the HMM. To incorporate spatial heterogeneity in the mutation rate along the genome, we modulate $\hat{\theta}_0$ by drawing scaling factors from a discretized Gamma distribution with mean equal to 1. The parameter shaping this prior distribution ($\alpha_\theta = \beta_\theta$) is estimated by maximum likelihood (via the forward HMM algorithm) together with other parameters of the model (using the Powell optimization procedure (Powell, 1964)). We model the changes in mutation rate along the genome as a Markov process with a single parameter δ_θ, the transition probability between any class of mutation rate, which is independent of the genealogical process. The justification for the Markov model is that sites in close proximity are expected to have similar mutation rates. For example, as is the case when the efficiency of the replication machinery decreases with increasing distance from the start of the replication fork (Francioli et al., 2015). Of note, Felsenstein & Churchill (1996) used a similar approach to model substitution rate variation across sites in a phylogenetic model. Let $n^{(\tau)}$ be the number of discretized τ intervals, and $n^{(\theta)}$ be the number of discretized categories of the prior distribution of scaling factors of θ. The ensuing Markov-modulated HMM has $n = n^{(\tau)} \times n^{(\theta)}$ hidden states. The transition matrix for spatial variation in θ is:

$$Q_\theta = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1n^{(\theta)}} \\ P_{21} & P_{22} & \cdots & P_{2n^{(\theta)}} \\ \vdots & \vdots & \ddots & \vdots \\ P_{n^{(\theta)}1} & P_{n^{(\theta)}2} & \cdots & P_{(n^{(\theta)}n^{(\theta)})} \end{bmatrix} = \begin{bmatrix} 1-\delta & \frac{\delta_\theta}{n^{(\theta)}-1} & \cdots & \frac{\delta_\theta}{n^{(\theta)}-1} \\ \frac{\delta_\theta}{n^{(\theta)}-1} & 1-\delta_\theta & \cdots & \frac{\delta_\theta}{n^{(\theta)}-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\delta_\theta}{n^{(\theta)}-1} & \frac{\delta_\theta}{n^{(\theta)}-1} & \cdots & 1-\delta_\theta \end{bmatrix}$$

(1)

where δ_θ is the aforementioned auto-correlation parameter. The resulting process is a combination of the SMC and the mutation Markov model, so that its transition probabilities are functions of the parameters from both processes, that is, the coalescence rates (parameterized by splines, similarly to (Terhorst et al., 2017)), δ_θ and the global recombination rate ρ (Barroso et al., 2019). The forward recursion for this model evaluated at genomic position *i* can be spelled out as:

$$F_i(\tau_t, \theta_m) = \left(\sum_{k=1}^{n^{(\theta)}} \left(\sum_{j=1}^{n^{(\tau)}} F_{i-1}\left(\tau_j, \theta_k\right) \cdot P\left(\tau_j \to \tau_t\right) \cdot P(\theta_k \to \theta_m)\right)\right) \cdot P(\tau_t \to S_i|\theta_m)$$

(2)

where θ_m is the product of $\hat{\theta}_0$ and the value of the *m*-th discretized category drawn from its prior Gamma distribution. The emission probability of binary state S_i depends on the height of the *t*-th genealogy and the focal mutation rate θ_m. More specifically, the emission probabilities of θ-iSMC are $P(0|\tau_t, \theta_m) = e^{(-\theta_m \times \tau_t)}$, and $P(1|\tau_t, \theta_m) = 1 - e^{(-\theta_m \times \tau_t)}$. Thus, the forward recursion integrates over all $n^{(\theta)}$ categories of θ and over all $n^{(\tau)}$ intervals of τ, for all sites in the genome. In the double-modulated model (ρ-θ-iSMC), where both mutation and recombination are allowed to vary along the genome, this integration is performed over θ, τ as well as ρ (giving a total of $n^{(\tau)} \times n^{(\theta)} \times n^{(\rho)}$ hidden states, Figure 1). Since spatial variation in ρ contributes to the transition probability between genealogies, the complete forward recursion is now given by:

$$F_i(\tau_t, \theta_m, \rho_r) = \left(\sum_{l=1}^{n^{(\theta)}} \left(\sum_{k=1}^{n^{(\rho)}} \left(\sum_{j=1}^{n^{(\tau)}} F_{i-1}\left(\tau_j, \rho_k, \theta_l\right) \cdot P\left(\tau_j \to \tau_t|\rho_k\right) \cdot P(\theta_l \to \theta_m) \cdot P(\rho_k \to \rho_r)\right)\right)\right) \cdot P(\tau_t \to S_i|\theta_m)$$

(3)

The full ρ-θ-iSMC model remains parsimonious, being characterized by a total of 11 parameters, namely, $\hat{\rho}_0$, $\hat{\theta}_0$, α_θ, α_ρ, δ_θ, δ_ρ plus five parameters describing constrained cubic splines that embody the demographic curve over time (Barroso et al., 2019). (Such parsimony is afforded by the structure of the Markov-modulated HMM which readily leverages physical linkage among sites in the same chromosome to fit distributions of recombination rates, mutation rates and TMRCA that are shared throughout the genome, even if site-specific realizations of these values may differ.) Running its forward recursion independently on each pair of genomes gives the composite likelihood of the model. After parameter optimization, we seek to reconstruct single-nucleotide landscapes (ρ, θ or τ) for each diploid separately. We first compute the posterior probability of each hidden state for every site *i* in the diploid genomes using regular HMM procedures (Durbin et al., 1998). Afterward, since in ρ-θ-iSMC the hidden states are triplets (Figure 1), computing the posterior average of each landscape of interest amounts to first marginalizing the probability distribution of its categories and then using it to weight the corresponding discretized values (Barroso et al., 2019). Let $\widehat{M}$ be the inferred discretized Gamma distribution shaping mutation rate variation, and $\hat{\theta}_l$ be the product of the estimated genome-wide average mutation rate $\hat{\theta}_0$ and m_l, the value of $\widehat{M}$ inside category *l*. Similarly, let $\hat{R}$ be the inferred discretized Gamma distribution shaping recombination rate variation, and $\hat{\rho}_k$ be the product of the estimated genome-wide average recombination rate $\hat{\rho}_0$ and r_k, the value of $\hat{R}$ inside category *k*. Then the posterior average $\hat{\theta}$ at position *i* is given by:

$$\hat{\theta}_l = \hat{\theta}_0 \cdot \sum_{l=1}^{n^{(\theta)}} m_l \cdot \left(\sum_{k=1}^{n^{(\rho)}} \sum_{j=1}^{n^{(\tau)}} P_i\left(\theta_l, \rho_k, \tau_j\right)\right)$$

(4)

where $P_i(\theta_l, \rho_k, \tau_j)$ is the probability of the triplet $\{\theta_l, \rho_k, \tau_j\}$ (which denotes a unique hidden state of the model) underlying the *i*-th site of the genome. Likewise, the posterior average $\hat{\rho}$ at position *i* is given by:

$$\hat{\rho}_l = \hat{\rho}_0 \cdot \sum_{k=1}^{n^{(\rho)}} r_k \cdot \left(\sum_{l=1}^{n^{(\theta)}} \sum_{j=1}^{n^{(\tau)}} P_i\left(\theta_l, \rho_k, \tau_j\right) \right)$$

(5)

Finally, the posterior average $\hat{\tau}$ at position *i* is presented in units of $4 \times Ne$ generations and obtained with:

$$\hat{\tau}_l = \sum_{j=1}^{n^{(\tau)}} \hat{\tau}_j \cdot \left(\sum_{k=1}^{n^{(\rho)}} \sum_{l=1}^{n^{(\theta)}} P_i\left(\theta_l, \rho_k, \tau_j\right) \right)$$

(6)

For each diploid, we can then bin the inferred single-nucleotide landscapes into non-overlapping windows of length *L* by averaging our site-specific estimates over all sites within each window. A consensus map of the population is obtained by further averaging over all *n* individual (binned) maps in our sample, i.e.:

$$\hat{\theta}_{pop}^{L} = \frac{1}{(n \times L)} \sum_{j=1}^{n} \sum_{i=1}^{L} \hat{\theta}_{i,j}$$

(7)

is our estimate of the consensus mutation rate in a single genomic window of length *L*, where *n* is the number of pairs of genomes analyzed by iSMC, and likewise for ρ and τ:

$$\hat{\rho}_{pop}^{L} = \frac{1}{(n \times L)} \sum_{j=1}^{n} \sum_{i=1}^{L} \hat{\rho}_{i,j}$$

(8)

$$\hat{\tau}_{pop}^{L} = \frac{1}{(n \times L)} \sum_{j=1}^{n} \sum_{i=1}^{L} \hat{\tau}_{i,j}$$

(9)

We finally note that the auto-correlation parameters δ_θ and δ_ρ represent the probabilities of switching mutation and recombination rates between adjacent sites, averaged along the genome. That is, although we include two layers of complexity in comparison to the original SMC models, we assume here that such transition probabilities are themselves spatially homogeneous. In reality, genomic regions may differ in the rate of change between local mutation and recombination rates. Nevertheless, in practice, the reconstruction of mutation and recombination maps with posterior decoding should be somewhat robust to this model mis-specification.

Simulation study

Using SCRM (Staab et al., 2015), we simulated 10 haploid sequences of length 30 Mb with parameters based on those inferred from ρ-θ-iSMC in *D. melanogaster* (see Results): θ = 0.0112; ρ = 0.036; α_θ (continuous Gamma distribution used as mutation rate prior) = 3.0; α_ρ (continuous Gamma distribution used as recombination rate prior) = 1.0; δ_θ (mutation rate transition probability) = 1e-05; δ_ρ (recombination rate transition probability) = 1e-04. Note that such transition probabilities lead to landscapes where blocks of constant mutation and recombination span, on average, 10 kb and 100 kb, respectively, with stochastic variation coming from the geometric distributions used to model them. Supplemental Figure S1 displays a sketch of the smoothed demographic history used in the coalescent simulations (see Results). Figures 4 and 5 display the mean R^2 value of the ANOVA performed on the inferred landscapes from 10 simulated replicates (see Results), but the standard deviation of these estimates are very small, and confidence intervals were, therefore, omitted. Data leading to Figure 5 was also simulated with SCRM, with parameters described in the Results section.

Next, we used SLiM 3.00 (Haller & Messer, 2018) to simulate the genealogy of a chromosome undergoing purifying selection, using *D. melanogaster*'s chromosome 2L as a template. The simulated region was 23.51 Mb long, and we used Comeron's recombination map in 100 kb windows (Comeron et al., 2012). We used Ensembl (Cunningham et al., 2022) release 103 gene annotations for *D. melanogaster* and extracted all exons coordinates, merging overlapping exons. Forward simulations were conducted using SLiM, with only deleterious mutations in exons being modeled. The fitness effect of mutations was drawn from a negative gamma distribution with a shape of 1.0 and a mean of -5/10,000. The population size was kept constant and equal to 10,000 and the population evolved for 700,000 generations. To compensate for the low population size, we scaled the mutation and recombination rates by a factor of 10 to result in a scenario closer to the *D. melanogaster* demography. The deleterious mutation rate was set to 1e-7 bp^{-1} along the genome. Ten replicates were generated and saved as tree sequences (Kelleher et al., 2018), which were then further processed by the 'pyslim' python module to run a recapitation procedure to ensure that all lineages coalesced into a single root at all genome positions. Ten genomes were then sampled uniformly at random and the underlying tree sequence exported. Finally, 'msprime' (Kelleher et al., 2016) was used to add neutral mutations to the tree sequence and save the resulting sequence alignments. A random mutation rate map was generated by sampling relative rates from a Gamma distribution with mean equal to 1.0 and with a shape parameter equal to 2.5, in segments with lengths drawn from a geometric distribution with mean equal to 100 kb. The resulting mutation relative rate map was then scaled by the genome average mutation rate of 1e-7 bp^{-1}.

Analyses of *Drosophila* data

Model fitting and posterior decoding by ρ-θ-iSMC in *D. melanogaster* data were performed using a hidden-states configuration of 30 τ intervals, five ρ categories and five θ categories. We used publicly available data – haplotypes ZI103, ZI117, ZI161, ZI170, ZI179, ZI191, ZI129, ZI138, ZI198 and ZI206 coming from the Zambia population in the Drosophila Population Genomics Project Phase 3 (Lack et al., 2015). Note that the following filters have been previously applied to these data by the original authors: A) heterozygous regions (maintained in the inbred individuals by selection due to recessive lethal alleles); B) three bp around called in-dels; C) long identity-by-descent stretches between genomes from the same location; as well as D) segments showing evidence of recent admixture (from outside Africa back into Africa) were all masked (turned to 'N' in the FASTA files). We assigned gaps and masked nucleotides in these FASTA sequences as "missing" data (encoded by the observed state '2' within iSMC, for which all hidden states have emission probability equal to 1.0 (Li & Durbin, 2011)). To optimize computational time, ρ-θ-iSMC was first fitted to chromosome 2L only. Maximum likelihood estimates from this model were then used to perform posterior decoding on all other autosomes. Prior to fitting the linear models, for each scale in which the iSMC-inferred landscapes were binned (50 kb, 200 kb and 1 Mb), we filtered out windows with more than 10% missing data in the resulting maps. Genomic coordinates for coding sequences and their summary statistics (π_N, and π_S) were taken from (Moutinho et al., 2019).

Linear modeling

Linear models implementing our causal model of diversity (Figure 3) were built based on genomic maps of 50 kb, 200 kb and 1 Mb resolution. It is worth reiterating that the binning of the single-nucleotide landscapes happens after optimization by the HMM such that it does not influence model complexity (as detailed in the model description, the 11 iSMC parameters are jointly estimated for the entire dataset, i.e., the model is aware of all individual sites in the sequences during optimization). When building linear models from real data, we first fitted GLS models independently to each autosome arm (2L, 2R, 3L, 3R), correcting for both auto-correlation of and heteroscedasticity of the residuals. After using Bonferroni correction for multiple testing, we observed (across the autosome arms and for different window sizes) significant and positive effects of θ and τ on π, whereas the effect of ρ was only significant for chromosome 3L at the 200 kb scale, and the interaction between θ and τ is positive and significant except for arms 2R and 3L at the 1 Mb scale (Supplemental Tables S6, S7, S8). Since the trends in coefficients are overall consistent, we pulled the autosome arms and in the Results section we present linear models fitted to the entire genome, for ease of exposition. Because we cannot rely on the GLS to partition the variance explained by each variable using type II ANOVA, we used OLS models to compute R^2 and restricted the GLS to assess the sign and significance of variables. We standardized all explanatory variables (subtracted the mean then divided by the standard deviation) before fitting the regression models to aid in both computation of variance inflation factors and interpretation of the coefficients.

Results

The sequentially Markov coalescent with heterogeneous mutation and recombination

The sequentially Markovian Coalescent (SMC) frames the genealogical process as unfolding spatially along the genome (Marjoram & Wall, 2006; McVean & Cardin, 2005; Wiuf & Hein, 1999). Its first implementation as an inference tool derives the transition probabilities of genealogies between adjacent sites as a function of the historical variation in N_e (i.e., demographic history) and the genome-wide average scaled recombination rate $\rho = 4 \times Ne \times r$ (Li & Durbin, 2011). Model fitting is achieved by casting the SMC as a hidden Markov model (HMM) (Dutheil, 2017) and letting the emission probabilities be functions of the underlying Time to the Most Recent Common Ancestor (TMRCA, τ) and the scaled mutation rate $\theta = 4 \times Ne \times \mu$ (see Methods). The SMC has proven to be quite flexible and serves as the theoretical basis for several models of demographic inference (see Spence et al. (2018) for a review, and Sellinger et al. (2020) for another compelling, more recent development). We have previously extended this process to account for the variation of ρ along the genome, thereby allowing for a heterogeneous frequency of transitions between local genealogies in different parts of the genome (Barroso et al., 2019). In this more general process called iSMC, recombination rate heterogeneity is captured by an auto-correlation parameter, δ_ρ, where the localized values of ρ are taken from a discrete distribution and the transition between recombination rates along the genome follows a first-order Markov process.

In the general case, the iSMC process is a Markov-modulated Markov process that can be cast as an HMM where the hidden states are *n*-tuples storing all combinations of genealogies and discretized values of each parameter that is allowed to vary along the genome (Dutheil, 2021). If one such parameter contributes to either the transition or emission probabilities of the HMM, then the hyper-parameters that shape its prior distribution can be optimized, e.g. by maximum likelihood (see Methods). In the iSMC with heterogeneous recombination (ρ-iSMC) the hidden states are pairs of genealogies and recombination rates (Barroso et al., 2019). Here, we extend this model by allowing the mutation rate to also vary along the genome (Figure 1), following an independent Markov process, i.e., letting the hidden states of the HMM be {θ-category, ρ-category, genealogy} triplets. The signal that spatial variation in ρ and θ leaves on the distribution of SNPs is discernible because their contributions to the likelihood are orthogonal: the recombination and mutation rates affect the transition and emission probabilities of the forward HMM algorithm, respectively. Parameter optimization and subsequent posterior decoding is performed as in Barroso et al. (2019). Under strict neutrality (which results in N_e being homogeneous along the genome (Charlesworth, 2009)), the inferred θ landscape reflects the landscape of *de novo* mutations (μ). iSMC can, therefore, be used to infer genome-wide variation in mutation rates with single-nucleotide resolution and statistical noise is reduced by averaging the posterior estimates of θ within larger genomic windows (see Methods).

In order to increase power, we further extend iSMC to accommodate multiple haploid genomes. In this augmented model, input genomes are combined in pairs such that the underlying genealogies have a trivial topology reduced to their τ (Figure 1). Although under Kingman's Coalescent (Kingman, 1982) the genealogies of multiple pairs of genomes are not independent, we approximate and compute the composite log-likelihood of the entire dataset by summing over such "diploid" log-likelihoods, similarly to MSMC2 (Malaspinas et al., 2016). Furthermore, iSMC enforces all diploids to share their prior distributions of τ, ρ and θ so that multiple sequences provide aggregate information to our parameter inference during model fitting; it does not, however, explicitly enforce that they have identical genomic landscapes upon posterior decoding. Rather, iSMC uses posterior probabilities to reconstruct recombination and mutation maps separately for each diploid.

Especially at the single-nucleotide level, accuracy of the inferred posterior landscapes is limited by the large stochasticity of the coalescent (Hein et al., 2004). The combination of genealogical and mutational variance leads to differences among the posterior landscapes of θ and ρ inferred from each diploid because it creates departures from the expected number of SNPs along pairs of genomes (hence variation in the amount of information diploids bear, in different regions of the genome, about ancestral processes such as mutation and recombination). To reduce noise from the individual diploid estimates and obtain consensus landscapes of the whole sample, iSMC averages the posterior estimates of θ and ρ over all diploids, for each site in the genome (see Methods). On the other hand, differences in the τ landscapes among diploids primarily reflect the stochastic nature of the ancestral recombination graph along the genome, which has intrinsic value itself. We therefore average these diploid τ landscapes not to reduce estimation noise but to obtain a measure of drift in neutral simulations. Note, however, that the average τ of the sample within a genomic window also contains information about natural selection (Palamara et al., 2018) – a property we exploit in the analyses of *Drosophila* data.

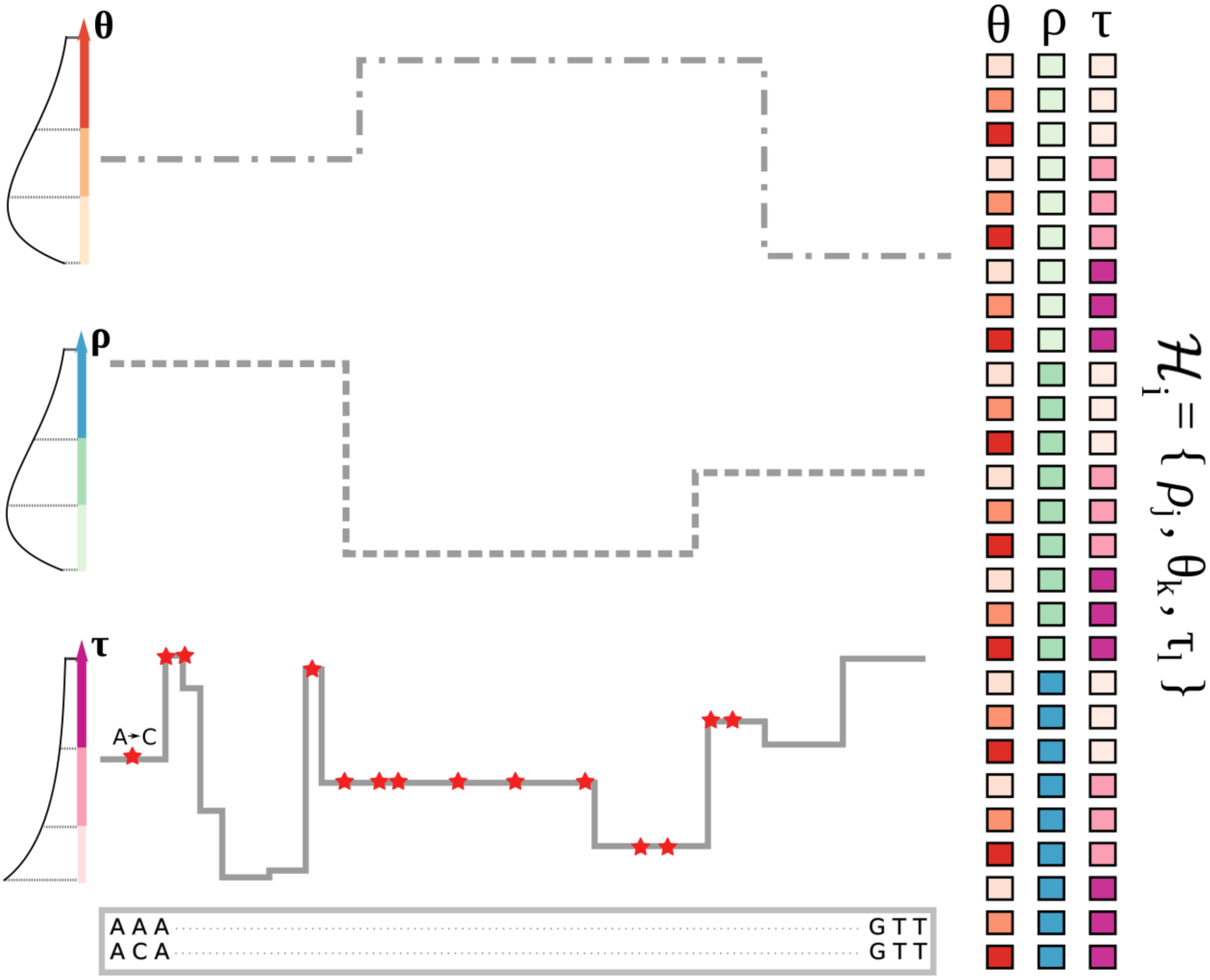

Figure 1 – Schematic representation of ρ-θ-iSMC for one pair of genomes. This cartoon model has three-time intervals, three recombination rate categories and three mutation rate categories. The genome-wide distribution of diversity depends on the mutation landscape (top) and on the τ landscape (bottom), which is modulated by the recombination landscape (middle). Discretized values of these distributions (left) are combined in triplets as the hidden states of our Hidden Markov Model (right).

Mutation rate variation impacts nucleotide diversity more than linked selection in *Drosophila*

We sought to quantify the determinants of genome-wide diversity in *D. melanogaster* using 10 haploid genome sequences from the Zambia population. To infer the genomic landscapes, we employed a ρ-θ-iSMC model with five mutation rate classes, five recombination rate classes and 30 coalescence time intervals, leading to 750 hidden states. We note that the number of classes and time intervals do not affect the number of estimated parameters, in particular because our implementation of the demographic model uses splines in place of the emblematic "skyline" model (Li & Durbin, 2011; Schiffels & Durbin, 2014) (see Methods). In general, finer discretization of these three distributions leads to more precise inference until a plateau is reached, as well as impacts the minimum and maximum values that the posterior estimates can take. However, the memory use and the likelihood computation time scale linearly and quadratically with the total number of hidden states, respectively. We selected 30 classes for the TMRCA and five classes for each rate distribution because this configuration provided a good trade-off between computational resources and accuracy during our testing phase. The total run-time for fitting the model with 750 hidden states to chromosome 2L of *D. melanogaster* was about 1 month on a high-performance cluster. Therefore, we proceeded in two steps: we first estimated model parameters on a subset of the data (chromosome arm 2L), and then used the fitted model to infer the landscape of mutation, recombination and TMRCA for all autosomes (see Methods). The justification for this approach is that the HMM posterior decoding is able to reconstruct chromosome-specific landscapes, even from identical prior distributions. At the same time, we have no *a priori* reason to believe that the shape of these distributions will differ substantially among autosomes. The similarity among the results obtained with each chromosome in the downstream analyses (see "Linear Modeling" sub-section within Methods) supports such intuition (Supplemental Tables S6, S7 and S8).

The iSMC parameters estimated from *D. melanogaster* suggest an exponential-like distribution of recombination rates ($\hat{\alpha}_\rho = \hat{\beta}_\rho$ ~1.03 for their Gamma distribution) whereas the inferred distribution of mutation rates is more tightly centered around the mean ($\hat{\alpha}_\theta = \hat{\beta}_\theta$ ~2.93 for their Gamma distribution). iSMC also inferred that the change in recombination rate across the genome was more frequent (auto-correlation parameter $\hat{\delta}_\rho$ ~0.9999, corresponding to a change of recombination rate on average every 10 kb) than the change in mutation rate (auto-correlation parameter $\hat{\delta}_\theta$ ~0.99999, corresponding to a change of mutation rate on average every 100 kb). This suggests that our model mostly captures large-scale rather than fine-scale variation in the mutation rate. Our inferred genome-wide average $\hat{\rho}$ (0.036) is in line with previous estimates (Chan et al., 2012), and the coalescence rates (which, in the context of this article, comprise a collection of nuisance parameters used to refine our estimates of τ, ρ and θ along the genome) suggest a ~4-fold bottleneck followed by recovery (Supplemental Figure S1). As an empirical validation of this new iSMC method, Spearman's rank correlations (hereafter referred to as Spearman's rho) between our inferred recombination map of chromosome 2L and Comeron's map based on experimental crosses (Comeron et al., 2012) are 0.594 at the 50 kb scale, 0.693 at the 200 kb scale and 0.865 at the 1 Mb scale (all p-values < 1e-5), higher than the correlations reported with previously published population genetic methods applied to *D. melanogaster* (Adrion et al., 2019; Barroso et al., 2019; Chan et al., 2012).

We used the parameters estimated from *D. melanogaster* to simulate 10 replicate datasets under a purely neutral scenario (see Methods). The aims of these simulations are two-fold: (1) to benchmark iSMC's accuracy in reconstructing the mutation landscape; and (2) to understand how ρ, θ and τ interact to influence diversity levels under neutrality, thereby providing a measure of contrast for the analyses of real data (where natural selection is present). Throughout this article, we analyze the determinants of nucleotide diversity at different scales by binning the landscapes of mutation, recombination and TMRCA into non-overlapping windows of 50 kb, 200 kb and 1 Mb. We first report strong correlations between inferred and simulated maps, ranging from 0.975 to 0.989 (Spearman's rho, Figure 2A, Supplemental Table S1), showcasing that our model is highly accurate under strict neutrality and when mutation rate varies along the genome in Markovian fashion.

We then used the raw genomic landscapes from these simulated (neutral) datasets to investigate how evolutionary mechanisms shape the distribution of nucleotide diversity along the genome, measured as π, the average per-site heterozygosity of the sample. The structure of our hypothesized causal model of diversity (solid lines in Figure 3) is rid of "backdoor paths" that would otherwise create spurious associations between recombination, mutation, TMRCA and nucleotide diversity (Pearl & Mackenzie, 2018). We could thus cast our causal model as an ordinary least squares regression (OLS) that seeks to explain π as a linear combination of the standardized variables ρ, θ and τ and statistical associations between our explanatory variables and the outcome variable π then represent causal relationships that merit scientific explanation. The justification for a linear model of π is that for sufficiently small genome-wide average diversity θ_0 (a requirement which is met in *D. melanogaster*, as $E[\theta_0]$ ~1e-2) the per-site heterozigosity $P(Heterozygous) = \pi = 1 - e^{(-\theta\times\tau)}$ can be well approximated by $\theta \times \tau$, the first term in the Taylor series expansion of $1 - e^{(-\theta\times\tau)}$. Since simulations grant direct access to the true genomic landscapes, then by definition the ensuing OLS models are free of estimation noise in the explanatory variables and serve as a ground truth assessment of how neutral evolutionary mechanisms influence nucleotide diversity. Because of the interplay between genealogical and mutational variance, we tested the improvement that including an interaction term between θ and τ brought to the fit of the linear models. In all replicates, we found that model selection using Akaike's information criterion favors a regression with an interaction term between the two variables that directly influence nucleotide diversity, $\pi_i = \beta_1 \cdot \tau_i + \beta_2 \cdot \theta_i + \beta_3 \cdot \rho_i + \beta_4 \cdot \theta_i{:}\,\tau_i + \epsilon_i$, over the simpler model $\pi_i = \beta_1 \cdot \tau_i + \beta_2 \cdot \theta_i + \beta_3 \cdot \rho_i + \epsilon_i$.

Fitting the regression model at the 50 kb, 200 kb and 1 Mb scales shows significant and positive effects of θ and τ, but not of ρ, on π (Supplemental Table S2, upper panel). This is expected since both deeper ancestry and higher mutation rate lead to increased nucleotide diversity and the influence of recombination rate on π is mediated by τ, thus disappearing due to its inclusion in the linear model. There is also a significant and positive effect of the interaction between θ and τ, highlighting the interplay between genealogical and mutational variance, where the effect of the mutation rate on diversity can only be fully manifested if ancestry is deep enough (reciprocally, ancestry can only be seen clearly if the local mutation rate is high enough). Moreover, the standardization that we employed on the explanatory variables prior to fitting the linear models (see Methods) allows us to evaluate their relative importance to the π distribution straight from the estimated coefficients. We observe that the linear coefficient of θ is ~6 times larger than the linear coefficient of τ at the 50 kb scale, ~11 times larger at the 200 kb scale and ~16 times larger at the 1 Mb scale (Supplemental Table S2, upper panel). Besides the linear coefficients, we further quantified the relative influence of mutation, drift and recombination to local diversity levels by partitioning the R^2 contributed by each explanatory variable with type II ANOVA. Consistently with the previous results, our estimates show that the θ landscape explains most of the variance in π in our simulations and that its contribution increases with the genomic scale (96.3% at 50 kb, 98.6% at 200 kb and 99.3% at 1 Mb Figure 4A). On the other hand, the contribution of the τ landscape decreases with the genomic scale (2.7% at 50 kb, 1% at 200 kb and 0.54% at 1 Mb). We propose that these trends stem from the minuscule scale of variation in τ (changing on average every 48.42 bp due to recombination events in our coalescent simulations, median = 19 bp), which smooth out more rapidly than does mutation variation when averaged within larger windows. Conversely, the broader scale of heterogeneity in θ (changing every 100 kb on average) makes it comparatively more relevant at larger window sizes. Strikingly, the total variance explained by the model is >99% at all scales, suggesting that these three landscapes are sufficient to describe the genome-wide distribution of diversity, as illustrated by our causal model (Figure 3).

To test whether we could recover such trends with the landscapes inferred by our HMM, we fitted the OLS models to the same genomic landscapes of nucleotide diversity except using the maps inferred by iSMC as explanatory variables (i.e., $\hat{\theta}$, $\hat{\tau}$ and $\hat{\rho}$ instead of the true, simulated ones: θ, τ and ρ). The sign and significance of the estimated OLS coefficients remained unchanged (Supplemental Table S2, middle panel), as do the ranking of their effect sizes, but in some replicates the residuals of the model were found to be correlated and/or with heterogeneous variance. As this violation of the OLS assumption could bias the estimates of the p-values of the linear coefficients, we also fitted Generalized Least Squares (GLS) models accounting for both deviations, which reassuringly produced coherent results (Supplemental Table S2, lower panel). Although co-linearity between $\hat{\theta}$ and $\hat{\tau}$ arises due to confounding in their estimation by iSMC, the variance inflation factors are always < 5, indicating that the coefficients are robust to this effect (Ferré, 2009). The trends in the linear coefficients obtained with iSMC-inferred landscapes are the same as those obtained with simulated (noise-free) landscapes, except that the effect of $\hat{\tau}$ is estimated to be larger than that of τ. Similarly, type II ANOVA using the inferred landscapes shows that the contribution of $\hat{\tau}$ is slightly higher than when using the true landscapes (5.1%, 2.9% and 1.4%, increasing window size) whereas the contribution of $\hat{\theta}$ is slightly lower (92.5%, 95.4% and 97.5%, increasing window size), but the variance explained by each variable closely agrees between the two cases (middle and right panels in Figure 4A). Therefore, we conclude that the joint-inference approach of iSMC can infer the genomic landscapes of τ, ρ and θ and that the linear regression representation of our causal model (Figure 3) is able to quantify their effect on the distribution of nucleotide diversity, π.

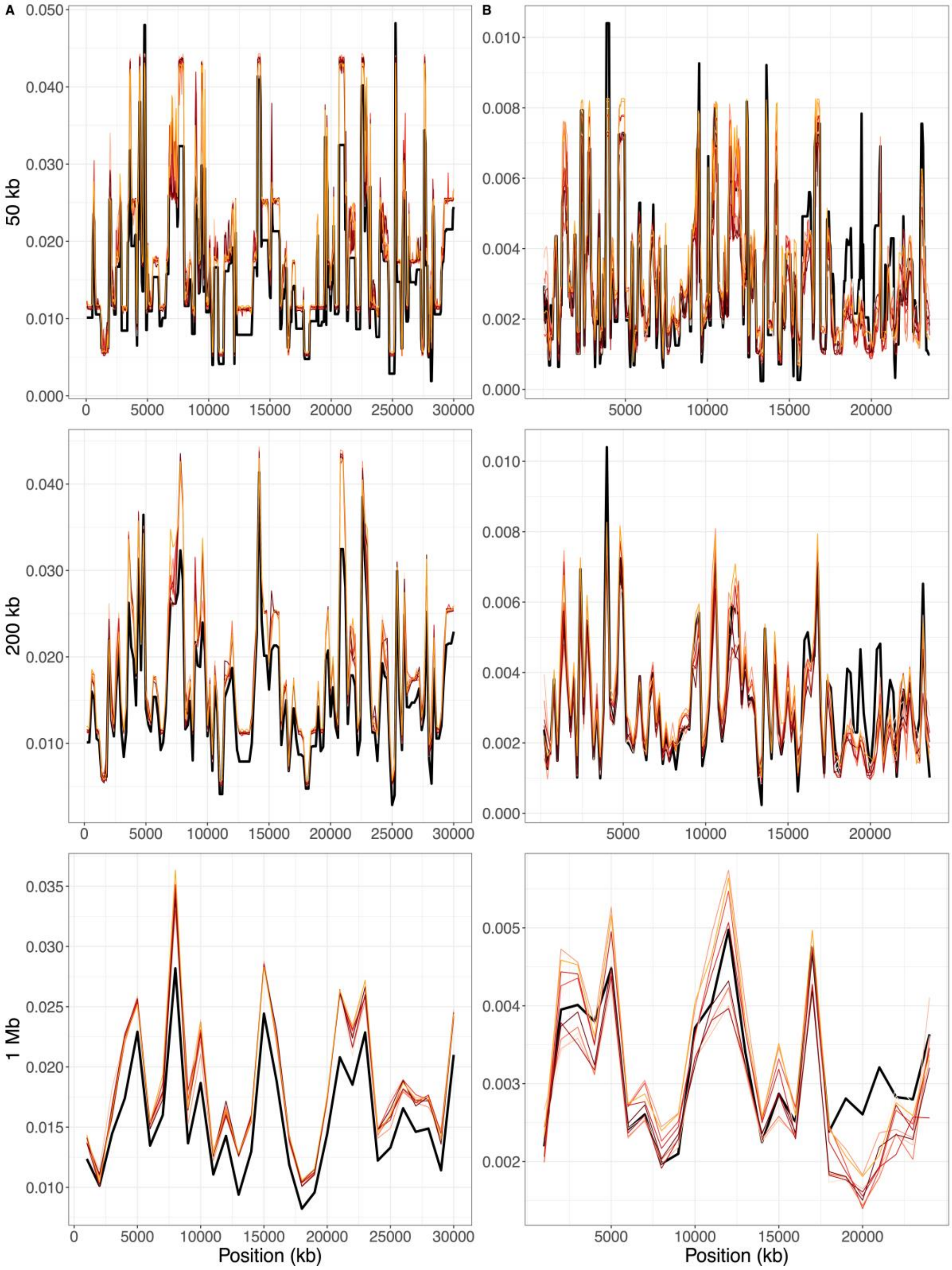

Figure 2 – iSMC recovers the mutation landscape in simulations. **A)** Coalescent simulations under neutrality. **B)** Simulations with background selection. In both cases, the simulated mutation landscape is shown by the thick black line whereas inferred landscapes are shown, for each replicate, by thin lines in shades of red. From top to bottom: 50 kb scale, 200 kb scale, 1 Mb scale.

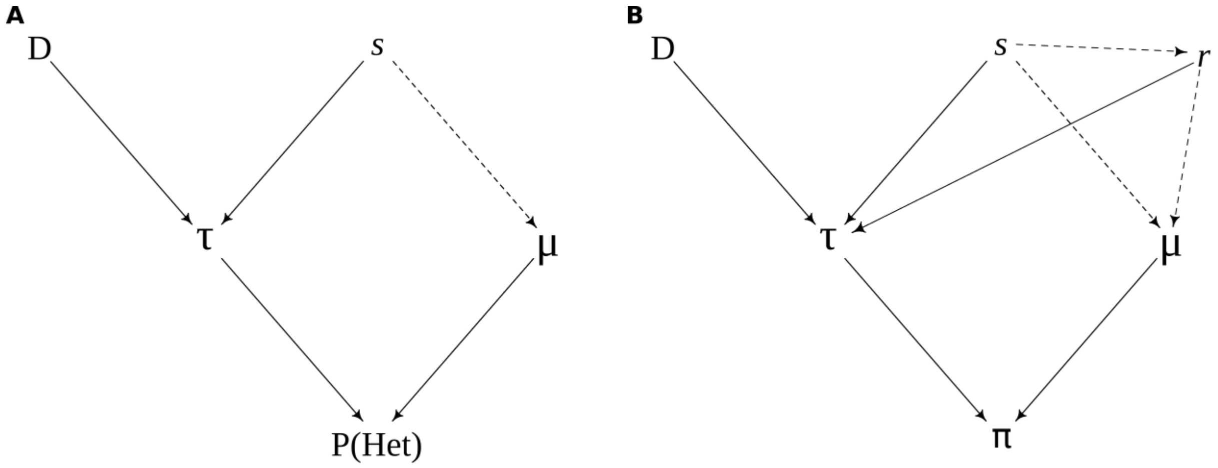

Figure 3 – Directed acyclic graphs depicting our abstract causal model for the determinants of genome-wide diversity. **A)** for a single, hypothetical nucleotide that is independent of any neighbors, its probability of being heterozygous is solely influenced by the local mutation rate (μ) and TMRCA (τ), which in turn is affected by drift (D) and selection (s). **B)** when contiguous sites are grouped into genomic windows, their correlated histories imply that the local recombination rate (r) plays a role in modulating both D and the breadth of linked selection via τ, which together with local μ influences π. Relationships that may be relevant in other model systems are shown by dashed lines (where selection affects μ and r through modifier genes and where recombination is mutagenic). Note that P(het) in **A** has exactly the same form as the emission probability of the HMM model, $P(1|\tau) = 1 - e^{(-\theta\times\tau)}$.

We finally employed the landscapes obtained with ρ-θ-iSMC to quantify the determinants of genome-wide diversity in *D. melanogaster*. In the following analyses, our interpretations of the OLS models assume that sequencing errors are unbiased with respect to the explanatory variables and that the population is broadly panmictic (or that geographic structure is implicitly accounted for by the TMRCA, e.g. (Beichman et al., 2018)). We also follow previous work suggesting that recombination is not mutagenic in this system (Begun et al., 2007; Castellano et al., 2016; McGaugh et al., 2012), thus we ignore this potential relationship. We used our inferred *D. melanogaster* maps to fit an OLS regression of the form $\pi_i = \beta_1 \cdot \hat{\tau}_l + \beta_2 \cdot \hat{\theta}_l + \beta_3 \cdot \hat{\rho}_l + \beta_4 \cdot \hat{\theta}_l{:}\hat{\tau}_l + \epsilon_i$. As in our simulations, the regression model shows positive effects of both $\hat{\theta}$ and $\hat{\tau}$, but not of $\hat{\rho}$, on π across all scales (Table 1). Likewise, a GLS model correcting for the identified auto-correlation of and heteroscedasticity of the residuals yields the same trends, and its variance inflation factors are < 5, indicating that the estimated coefficients are robust to co-linearity (Ferré, 2009). Showcasing its dominant impact on π in the fruit fly, the linear coefficient of $\hat{\theta}$ is between three and four times larger than that of $\hat{\tau}$, a trend that is akin to that obtained with inferred maps in the coalescent simulations. Moreover, partitioning of variance shows a small contribution of $\hat{\tau}$ that decreases with increasing genomic scale (5.9% at 50 kb, 2.1% at 200 kb and 2.1% at 1 Mb) whereas the opposite applies to $\hat{\theta}$ (91.7% at 50 kb, 96.7% at 200 kb and 96.8% at 1 Mb, left panel in Figure 4A). Our linear model explains >99% of the variation in π along *D. melanogaster* autosomes, and the effects of our inferred landscapes on diversity are remarkably close to those from our neutral simulations (Figure 4A), suggesting that iSMC is robust to the occurrence of selection in this system. Unlike neutral simulations; however, the simple correlation test between $\hat{\rho}$ and π ends up positive and significant in *D. melanogaster* data, at least at smaller scales (Spearman's rho = 0.20, p-value = 2e-13 at the 50 kb scale; Spearman's rho = 0.15, p-value = 0.0025 at the 200 kb scale; Spearman's rho = 0.20, p-value = 0.07 at the 1 Mb scale), recapitulating the classic result of Begun & Aquadro (1992) and indicating the presence of linked selection. We also found a positive correlation between $\hat{\rho}$ and $\hat{\tau}$ (Spearman's rho = 0.48, p-value < 2.2e-16 at 50 kb; Spearman's rho = 0.45, p-value < 2.2e-16 at 200 kb; Spearman's rho = 0.48, p-value < 2.2e-16 at 1 Mb), once again contrasting the results under neutrality and suggesting that the effect of linked selection is indeed captured by the distribution of genealogies and modulated by the recombination rate (Cutter & Payseur, 2013). Although τ is primarily influenced by demography in SMC-based models (by means of a Coalescent prior taming the transition probabilities of the HMM (Li & Durbin, 2011; Schiffels & Durbin, 2014), it has also been demonstrated to carry the signature of selection due to local changes in coalescence rates that have been interpreted as spatial variation in N_e (Palamara et al., 2018; Zeng & Charlesworth, 2011). Shortly, Palamara's ASMC method reconstructs the TMRCA landscape of several pairs of genomes and interprets recurrent (shallow) outliers in the $\hat{\tau}$ distribution as the outcome of linked selection (i.e., regions where pair of genomes consistently coalesce faster than expected under neutrality). We tested the sensitivity of our regression framework to this effect by a fitting linear model without $\hat{\tau}$ as an explanatory variable, $\pi_i = \beta_1 \cdot \hat{\theta}_l + \beta_2 \cdot \hat{\rho}_l + \epsilon_i$, hypothesizing that in the absence of its mediator the recombination rate would show a significant and positive effect on diversity. Indeed, this is what we found at all genomic scales (Table 1, Model 3), corroborating our interpretation of the causal relationships in the presence of selection (Figure 3B), from which the direct correlation between $\hat{\rho}$ and π, often reported in the literature, is a special case. In summary, our results show that recombination shapes diversity via the τ distribution and linked selection, but that in *D. melanogaster*, the impact of genetic hitchhiking on the diversity landscape is smaller than that of mutation rate variation.

Table 1 – Estimates from linear regression models fitted to the distribution of nucleotide diversity along *Drosophila melanogaster* genomes. Vertical panels show results according to genomic window size whereas horizontal panels show results according to the structure of the linear model. OLS = Ordinary Least Squares; GLS = Generalized Least Squares; VIF = Variance Inflation Factor.

			50 kb			**200 kb**			**1 Mb**		
Model	Type	Variable	Coefficient	p-value	VIF	Coefficient	p-value	VIF	Coefficient	p-value	VIF
1	OLS	θ	0.0027	<2.2e-16	1.3	0.0026	<2.2e-16	1.7	0.0025	<2.2e-16	2.4
		τ	0.0010	<2.2e-16	2.5	0.0008	<2.2e-16	6.0	0.0006	<2.2e-16	13.0
		ρ	0.00004	0.0788	1.5	0.00001	0.2110	1.7	0.0000003	0.0350	1.8
		θ:τ	0.0003	<2.2e-16	1.6	0.0002	<2.2e-16	3.7	0.0001	<1e-4	7.7
2	GLS	θ	0.0027	<1e-4	1.2	0.0026	<1e-4	1.4	0.0024	<1e-4	1.6
		τ	0.0010	<1e-4	1.9	0.0008	<1e-4	3.7	0.0007	<1e-4	1.8
		ρ	0.000004	0.3897	1.4	0.000005	0.5720	1.5	0.000003	0.7980	1.8
		θ:τ	0.0003	<1e-4	1.3	0.0002	<1e-4	2.5	0.0001	0.0003	1.3
3	GLS	θ	0.0030	<1e-4	1.0	0.0029	<1e-4	1.0	0.0027	<1e-4	1.1
		ρ	0.00040	<1e-4	1.0	0.0003	<1e-4	1.0	0.0002	<1e-4	1.1

To investigate the signature of selection, we analyzed the relationship between the local mutation rate and the levels of synonymous (π_S) and non-synonymous (π_N) diversity across *D. melanogaster* genes (see Methods). We computed these summary statistics across exons and matched their coordinates with our finest (50 kb-scale) genomic landscapes to increase resolution (i.e., to maximize variation in mutation and recombination rates among genes). We observed a stronger relationship between $\hat{\theta}$ and π_S (Spearman's rho = 0.68, 95% CI after 10,000 bootstrap replicates = [0.64, 0.72], partial correlation accounting for $\hat{\tau}$) than between $\hat{\theta}$ and π_N (Spearman's rho = 0.27, 95% CI after 10,000 bootstrap replicates = [0.22, 0.32], partial correlation accounting for $\hat{\tau}$) indicating that selection partially purges the excess of non-synonymous deleterious variants in genes with elevated mutation rate, whereas synonymous variants segregate more freely either because they are not directly affected by selection (but are still linked to selected sites) or because selection on codon usage (Lawrie et al., 2013; Machado et al., 2020) is not as strong as selection on protein function. Since synonymous variants are interdigitated with non-synonymous variants, the contrast between these correlation tests cannot be explained by a bias in iSMC's estimation of θ in functionally constrained regions of the genome. Furthermore, a correlation test between $\hat{\theta}$ and the proportion of exonic sites in the same 50 kb windows (Spearman's rho = -0.037, p-value = 0.19, partial correlation accounting for $\hat{\tau}$) fails to reveal such putative bias (see Discussion for a flip side view on this test). Conversely, we observed a negative and significant correlation between $\hat{\tau}$ and the proportion of exonic sites (Spearman's rho = -0.158, p-value = 2e-12, partial correlation accounting for $\hat{\theta}$), as expected since background selection should reduce the TMRCA more abruptly in densely constrained regions (Charlesworth, 2013; Palamara et al., 2018). We also fitted linear models considering only 50 kb windows with more than 20,000 coding sites. Once again, there were significant and positive effects of both $\hat{\theta}$ and $\hat{\tau}$, but not of $\hat{\rho}$, on π. Moreover, the mutation landscape remains the most important factor, explaining 93.2% of the distribution of diversity in gene-rich regions.

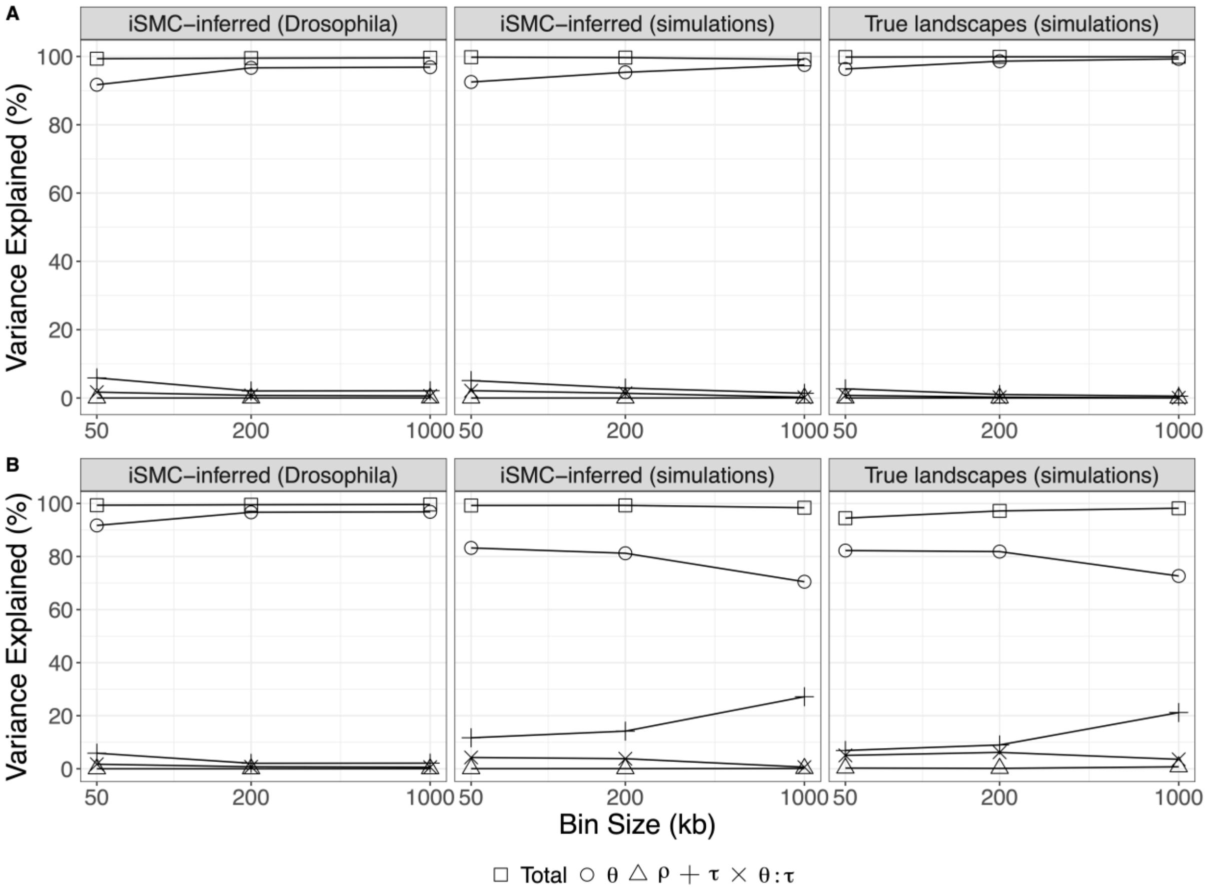

Figure 4 – Variance in the distribution of diversity explained by each genomic landscape. Partitioning of variance according to window size (x-axis, shown in $\log_{10}$ scale), using either simulated data (true landscapes: right panels; inferred landscapes: middle panels) or real *Drosophila* data (left panels). **A)** comparison between real *Drosophila* data and results from neutral simulations. **B)** comparison between real *Drosophila* data and results from simulations under background selection. In each panel, shapes represent explanatory variables in the linear model: θ (circles), ρ (triangles), τ (plus sign), θ:τ interaction (crosses) and the total variance explained by the model (squares) is the sum of the individual R^2. Each point represents the average R^2 over 10 replicates. Variation among replicates resulted in confidence intervals too small to be plotted.

Mutation rate variation shapes genome-wide diversity in neutral scenarios

Our analyses of *D. melanogaster* data and *D. melanogaster*-inspired simulations suggest that the mutation landscape is the main factor influencing levels of diversity along the genome. But are there scenarios where τ has a more pronounced effect on π? We addressed this question by exploring the parameter space of our neutral simulations. For fixed values of the long-term average population size (N_e = 100,000), the average mutation rate per site per generation (μ = 2e-09), the Gamma distribution of scaling factors of θ ($\alpha_\theta = \beta_\theta = 2.5$) and the Gamma distribution of scaling factors of ρ ($\alpha_\rho = \beta_\rho = 1.0$), we varied the demographic history (flat N_e; 10-fold bottleneck happening 0.5 coalescent time units ago), the average recombination rate per site per generation (*r* = 1e-08; 1e-09) and the fluctuations of the mutation landscape, where the realized lengths of genomic blocks of constant μ were drawn from geometric distributions with averages equal to 50 kb, 500 kb, or instead taken as a perfectly flat mutation landscape. We reasoned that the extent of the variation in τ along the genome compared to that of μ (equivalently, θ) should modulate their relative influence on π. We fitted OLS models to explain π using the true, simulated landscapes as explanatory variables, and computed their average R^2 over all replicates for each evolutionary scenario (Figure 5). The OLS models included an interaction term between θ and τ but its individual R^2 was excluded from the plots because it is overall low (~1%) and of no direct interest. We observed clear trends emerging from these simulated data. First, for a given demographic history and pattern of variation in the mutation rate, increasing *r* reduces the influence of τ on π. This happens because with high recombination rates the genealogies change more often along the genome, thus displaying more homogeneous maps when averaged within windows (50 kb, 200 kb, 1 Mb). Second, for a given *r* and pattern of variation in the mutation rate, τ has a larger impact on π in the bottleneck scenario compared to the scenario of constant population size. This happens because when N_e varies in time, the distribution of coalescence times becomes multi-modal (Hein et al., 2004) and therefore more heterogeneous along the genome. Third, for a given demographic history and fixed value of *r*, frequent changes in μ along the genome (on average every 50 kb) reduce its impact on π relative to rare changes in μ (on average every 500 kb). This happens because frequent changes in μ lead to more homogeneity along the genome, when averaged within the window sizes used in our analyses.

Finally, if the mutation landscape is flat, then, as expected, the variance explained by our linear model is entirely attributed to τ. Note that although in these neutral simulations τ varies along the genome as a result of genetic drift alone, it still has a non-negligible effect on the distribution of diversity in most scenarios (i.e., binning into large genomic windows does not flatten the TMRCA landscapes completely). This is in agreement with an observation that heterogeneous recombination rates lead to outliers in genome-wide F_{ST} scans, even under neutrality (Booker et al., 2020), which in turn happens because the recombination landscape enlarges the variance of the τ distribution by making the frequency of genealogy transitions a function of the local ρ (confirming the causal effect ρ → τ depicted in Figure 3B). From a practical standpoint, it means that drift should not be neglected as an explanation for the distribution of π, especially at narrow window sizes (≤ 10 kb). This is relevant because it is also at narrow window sizes that the effect of selection on diversity levels along the genome can be more easily confounded by the effect of drift, with extreme examples happening during population range expansions and especially in regions of low recombination (Schlichta et al., 2022).

More generally, our simulation study of neutral scenarios shows that the relative impacts of evolutionary mechanisms on π depend primarily on (1) the joint patterns of variation of ρ, τ and θ along the genome; and (2) the window size used in the analysis, because of averaging effects when building the genomic landscapes. In light of these results, the genome of *D. melanogaster* – with its high effective recombination rate, broad (as detectable by iSMC) pattern of variation in the mutation rate and high density of functional sites – seems to be particularly susceptible to the effect of the mutation landscape on its large-scale distribution of diversity. Yet, since the mutation landscape stood out as the most relevant factor in all of the explored (neutral) scenarios where it was allowed to vary (Figure 5), we predict that it is likely to shape genome-wide diversity patterns in other species as well.

iSMC can disentangle mutation rate variation from linked selection

Finally, we simulated 10 replicate datasets under a background selection model with genomic features partially mimicking those of *D. melanogaster* chromosome 2L (see Methods). Briefly, we included in these simulations the positions of exons from the Ensembl database (Cunningham et al., 2022) (whose non-synonymous mutations had selection coefficients drawn from a negative Gamma distribution), the (Comeron et al., 2012) recombination map of the fruit fly, and spatial variation in mutation rates using parameters estimated in our previous analyses. These datasets are not meant to precisely reproduce patterns of nucleotide diversity in real data – there are far too many biological processes not captured by the simulations –, but instead to assess iSMC's ability to disentangle the θ and τ landscapes when the latter is heavily distorted by linked selection. As such, we used a distribution of selection coefficients with a shape parameter equal to 1.0 (which contrasts with the range from ~0.3 to ~0.4 reported in the literature, e.g. Castellano et al. (2018)) in order to artificially exacerbate the effect of linked selection (see below). As before, we fitted a ρ-θ-iSMC model with five mutation rate classes, five recombination rate classes and 30 coalescence time intervals, and afterwards binned the simulated and inferred maps into windows of 50 kb, 200 kb and 1 Mb. We then assessed the accuracy of our framework in two ways: first, by computing Spearman's rho between simulated and iSMC-inferred landscapes; second, by contrasting the variance explained by each variable in the OLS regression (fitted with inferred versus true maps).

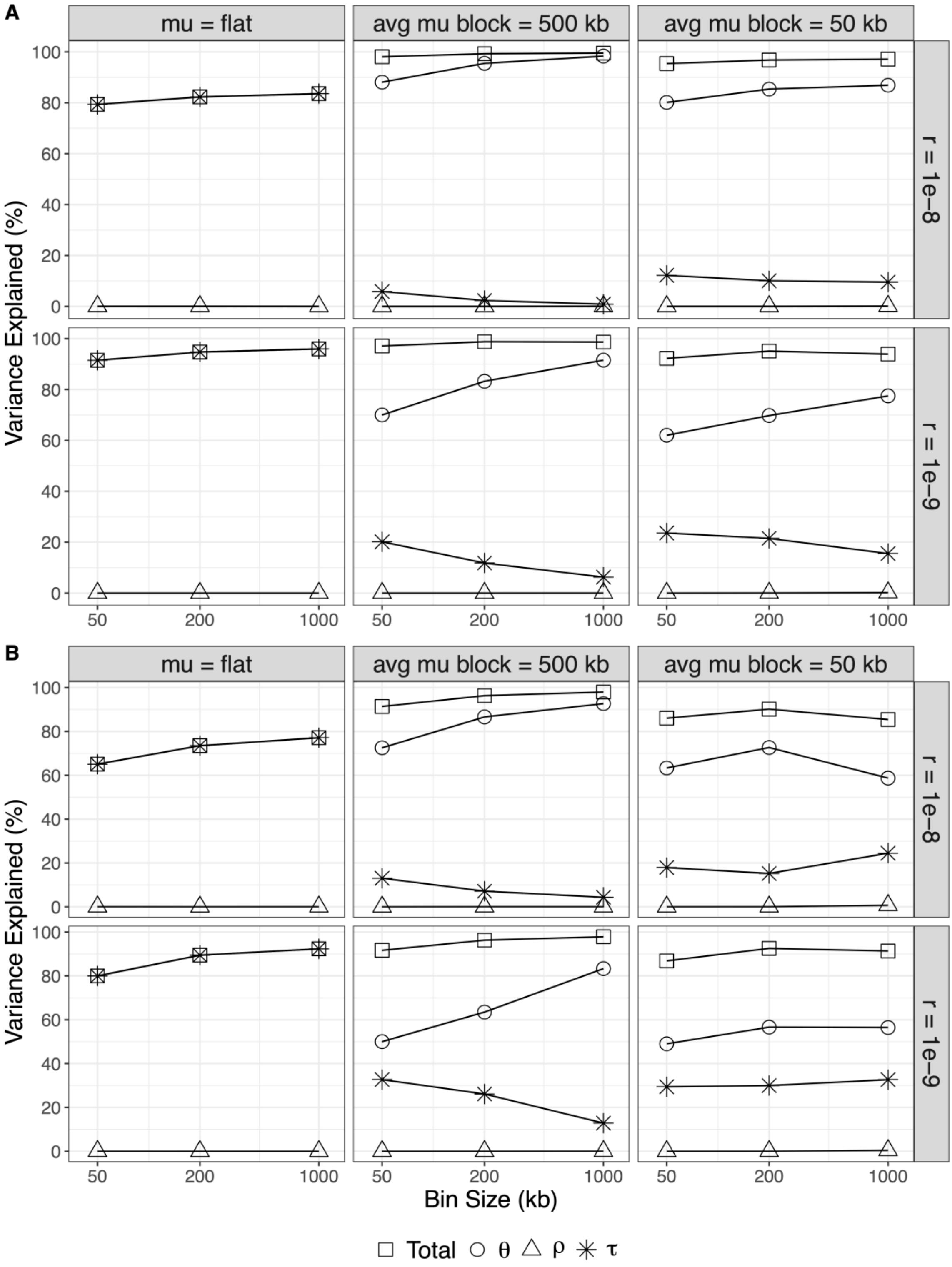

Figure 5 – Variance in the distribution of diversity explained by each genomic landscape (neutral simulation study). Partitioning of variance according to window size (x-axis, shown in $\log_{10}$ scale). **A)** Constant population size. **B)** Population bottleneck. Results are displayed according to parameters (rows = recombination rate, columns = scale of mutation rate variation). In each panel, point shapes represent explanatory variables in the linear model: θ (circles), ρ (triangles), τ (asterisks) and the total variance explained by the model (squares) is the sum of the individual R^2. Each point represents the average R^2 over 10 replicates, and variation among replicates resulted in confidence intervals too small to be plotted. All linear models were built using simulated (true) landscapes.

We report strong and positive correlations between simulated and inferred landscapes under such complex a scenario (Figure 6). As expected, the inherently fine-scale variation in the distribution of genealogies is the hardest to reconstruct: the Spearman's rho between the true TMRCA landscapes and the inferred ones ranges from 0.385 to 0.465 at 50 kb and increases substantially with window size (up to 0.787 at 1 Mb, Supplemental Table S3). In comparison, the correlation between the true and inferred mutation landscapes ranges from 0.751 (at 1 Mb) to 0.894 (at 50 kb, Supplemental Table S4). The recombination landscape is also well recovered, with correlation coefficients ranging from 0.830 (at 50 kb) to 0.963 (at 1 Mb, Supplemental Table S5). We postulate that iSMC's power to distinguish between the signal that θ and τ leave on sequence data stems exactly from the difference in scale at which they vary along the simulated chromosomes. Although linked selection can increase the correlation among genealogies around constrained sites (McVean, 2007), in most genomic regions the extent of such effect is still short in comparison to the scale of mutation rate variation, allowing their effects on the distribution of diversity to be teased apart. In summary, although under linked selection the accuracy of inferred mutation maps is lower than under strict neutrality (cf. Supplemental Table S1), it remains high enough to validate the robustness of our new model of mutation rate variation.

Given the high accuracy of iSMC in these challenging simulations, one would expect hefty resemblance in the linear models when using the inferred versus the true, noise-free landscapes. This is indeed what we found at all scales (middle and right panels in Figure 4B), suggesting that residual biases in iSMC due to linked selection do not carry over to the regression analyses noticeably. Moreover, there is a closer agreement of real *D. melanogaster* data with neutral simulations than with simulations of background selection. This is probably a consequence of the unrealistically strong background selection in the simulations, where the distribution of fitness effects we used has a high density of weakly deleterious mutations which segregate longer in the population, leading to a more localized and pronounced distortion of genealogies (Zeng & Charlesworth, 2011). This would also explain why the R^2 attributed to θ and τ respectively decrease and increase with window size (Figure 4B), a reverse relationship than observed in the real data and in neutral simulations (Figure 4A, Figure 5). These results are corroborated by the linear coefficients (Supplemental Table S9). In the presence of selection, the coefficient of θ decreases with window size and is distinctively closer to the coefficient of τ than under neutrality, a relationship that is reproduced when fitting the linear models with $\hat{\theta}$ and $\hat{\tau}$. Note also that under intense background selection the coefficient of ρ is generally larger than in the previous scenarios (and the corresponding p-values smaller), mis-matching real *D. melanogaster* data as well. We further inquired into these phenomena visually, by looking into the simulated landscapes, which provided critical insight into the interplay among micro-evolutionary mechanisms shaping diversity (Figure 7).

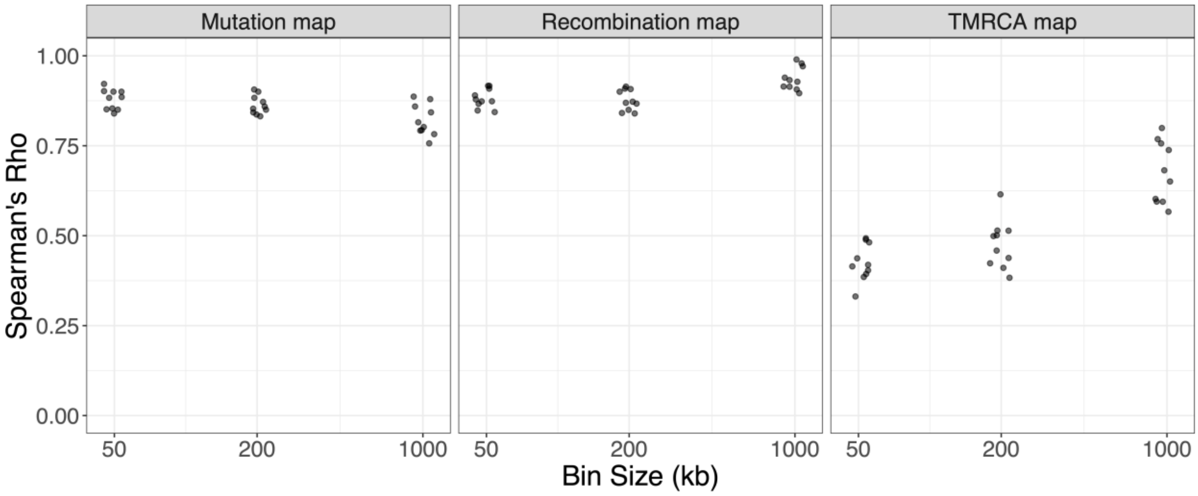

Figure 6 – Spearman correlations between inferred and simulated landscapes under background selection. All p-values are smaller than 1e-4.

Close inspection shows that only in regions of extremely reduced recombination (the left tip and right tail of the simulated chromosome) does linked selection introduce enough correlation among selected and neutral sites as to influence diversity to a larger degree than mutation rate variation, and this effect seemingly grows with window size. Otherwise, the distribution of π predominantly mirrors that of θ, endorsing our previous results. As a side note, Figure 7 lays out a rather enticing graphic of our linear models: they seek to represent π (the top row) as a linear combination of τ, ρ, and θ (the other three rows), plus the interaction τ:θ. From this perspective, it becomes apparent that the mutation landscape contributes the most to variation in diversity along the chromosome, even in such a conservative scenario where linked selection is artificially strong. Taken together, the results from these simulations provide compelling evidence that the high R^2 attributed to θ in *D. melanogaster* is a solid finding. First, it cannot be explained either by the increased noise in our inference of τ compared to θ (Figure 6), or by potential absorption of linked selection effects into $\hat{\theta}$, since in both of these cases we would not expect the close correspondence between OLS results fitted with inferred versus true maps (Figure 4b). Second, raw simulated data clearly demonstrate that the effect of linked selection can be overwhelmed by mutation rate variation (Figure 7). We conclude that the modeling framework illustrated in Figures 1 and 3 satisfactorily captures the essence of the genome-wide determinants of nucleotide diversity, and is likewise adequate to the study of *D. melanogaster.*

Discussion

The presence of mutation rate variation along the genome has been recognized for many years (some of the evidence in mammals reviewed over a decade ago by Hodgkinson & Eyre-Walker (2011)), although its implications have been largely overlooked by the population genetics literature. The contributions of the present work are not to simply recapitulate this phenomenon in *D. melanogaster*, but mainly to (1) present a novel statistical method that can infer such variation using population genetic data and (2) use this method to show that the mutation landscape has a lasting effect on nucleotide diversity patterns that can be quantitatively larger than that of natural selection. This awareness is long overdue, as the relative strengths of selection and drift in shaping genome-wide diversity have been debated for several decades (reviewed in Hey (1999); and, more recently, Kern & Hahn (2018) and Jensen et al. (2019)), with the influence of local mutation rate only recently brought to light (Castellano et al., 2018; Harpak et al., 2016; Smith et al., 2018). We were able to employ our extended iSMC model to jointly infer mutation, recombination and TMRCA landscapes and to use causal inference to estimate their impact on π along the genome. Our analyses revealed that these combined landscapes explain >99% of the distribution of diversity along the *D. melanogaster* genome; when looking into the detailed patterns, we found the footprints of linked selection, but the major driver of genome-wide diversity in this species seems to be the mutation landscape. Importantly, this conclusion holds whether we base the discussion on estimates of linear coefficients or on the proportion of variance explained.

These results do not imply that linked selection cannot extend beyond the 18.3% of the *D. melanogaster* genome that is exonic (Alexander et al., 2010), but rather that variation in the mutation rate is strong enough to contribute relatively more to the variation in π, in the genomic scales here employed (Figure 4). Our findings, however, sharply contrast with an estimate by Comeron that up to 70% of the distribution of diversity in *D. melanogaster* can be explained solely by background selection at the 100 kb scale (Comeron, 2014), where the author further argued that many regions of increased diversity may be experiencing balancing selection. Instead, we propose that mutation rate variation is responsible for most of these effects. We believe that such discrepancy can be mainly attributed to Comeron's 70% figure deriving from the (rank) correlation between π and B-value maps alone, without including other causal factors (like drift and local μ). The B-value represents the expected reduction of diversity due to selection against linked deleterious mutations (Charlesworth, 2013; Matheson & Masel, 2022; McVicker et al., 2009). This is equivalent to a scaling of the expected TMRCA between two (uniformly) random samples, which in our model is captured by $\hat{\tau}$. Indeed we find that despite $\hat{\tau}$ explaining little variance in diversity in the multiple regression setting, the simple correlations between $\hat{\tau}$ and π are of the same order as found by Comeron (Spearman's rho = 0.70, p-value < 2.2e-16 at the 50 kb scale; Spearman's rho = 0.66, p-value < 2.2e-16 at the 200 kb scale; Spearman's rho = 0.86, p-value < 2.2e-16 at the 1 Mb scale, cf. Table 1 in (Comeron, 2014)). The central point being that the linear models were able to reliably pinpoint θ as the main driver of π just because its effect was jointly estimated with those of τ and ρ. Taking a step back, it is also conceivable that selection is not only manifested as distortions in the distribution of genealogies, but also biases our estimate of the mutation landscape. We note, however, that there actually seems to be a

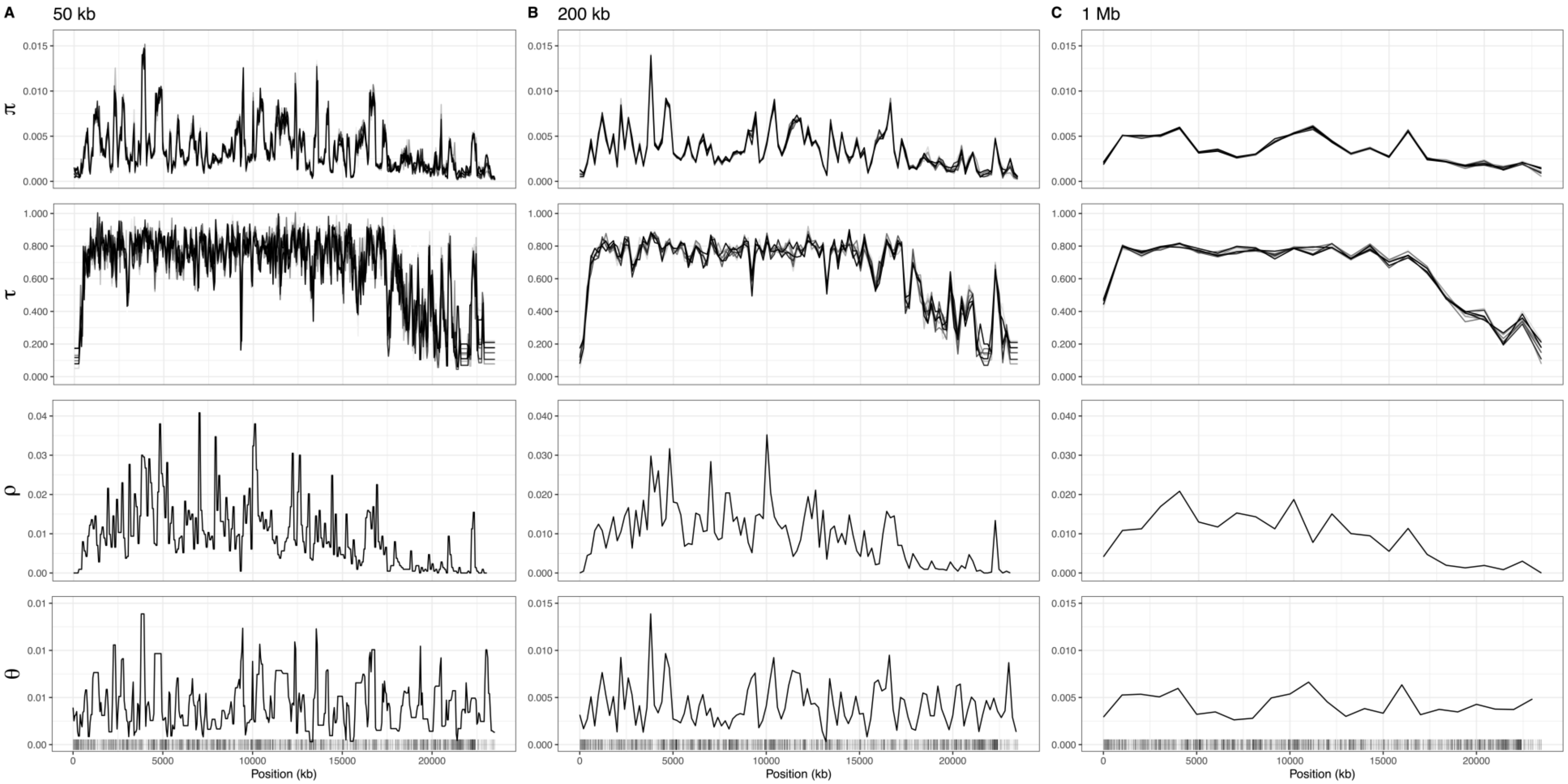

Figure 7 – Genomic landscapes simulated under background selection and their effect on the distribution of diversity. From top to bottom: observed nucleotide diversity (lines in shades of grey represent replicates), average TMRCA of the genealogies in units of generations (lines in shades of grey represent replicates), the recombination landscape (shared among replicates) and the mutation landscape (shared among replicates). All landscapes are binned into non-overlapping windows of 50 kb (**A**), 200 kb (**B**) and 1 Mb (**C**). Barcode at the bottom represents the density of sites under negative selection. These data were extracted straight from the simulations and are, therefore, free of estimation noise.

small overestimation of the importance of the TMRCA in our results (compare values obtained with true vs inferred landscapes in Figure 4 and Supplemental Tables S2, S9), which goes in the opposite direction to the presumed bias under linked selection. In this way our results appear to be conservative with respect to the discussions we submitted throughout this article. On top of that, based on the high similarity between real *D. melanogaster* data and our neutral simulations (Figure 4A) as well as on iSMC's robustness to the presence of linked selection (Figure 2B, Figure 4B, Figure 6), we argue that a bias induced by linked selection would likely be insufficient to overturn our conclusion of a major impact of mutation rate variation on the distribution of diversity.

We also note that selection should have a stronger impact on π when binning is performed at smaller genomic scales (≤10 kb, e.g., Figure 4 in Hudson & Kaplan (1995)), which we have not explored because of increased genealogical and mutational variance at such small window sizes. Besides Comeron, Elyashiv et al. (2016) also used patterns of nucleotide variation to fit models of linked selection along the fruit fly genome. Using substitution rates at synonymous sites as a proxy for local mutation rates, they employed their selection estimates to predict genome-wide diversity in windows from 1 kb to 1 Mb. Their models predict 44% (100 kb) and 76% (1 Mb) of the distribution of scaled nucleotide diversity in *D. melanogaster*. However, owing to the scaling that removes the effect of mutation rate variation, the percentages in the Elyashiv et. al study represent the part of variance explained by linked selection once the effect of mutation rate variation has been discarded (see also Murphy et al. (2022) for a similar and improved model). As such, the R^2 values they report quantify the goodness-of-fit of different models of selection (e.g., background selection alone vs background selection + selective sweeps) instead of the actual importance of linked selection to π, and are, therefore, not directly comparable with our estimates. Still, we note that the remaining variance in their models may be due to mutation rate variation not grasped by synonymous divergence – an imperfect proxy for μ, either because of selection on codon usage or because the mutation landscape has evolved since the divergence of the two species. (Along the same vein, the correlation between our 50 kb mutation maps and genome-wide divergence between *D. melanogaster* and *D. yakuba* is only moderate, Spearman's rho = 0.197, p-value = 3e-09.) The differences between our approaches to capture linked selection are also worth discussing. While Elyashiv et al. (2016) relied on elaborate models of selection that embody strong assumptions, we leaned on the spatial distribution of τ, similarly to Palamara et al. (2018). This heuristic renders our approach more parsimonious (11 parameters compared to 36 in the Elyashiv et. al model) and less susceptible to mis-specifications of the selection model, which could be commonplace (e.g., the presence of epistasis and/or fluctuating fitness effects over time). Developing an explicit model of spatial variation in N_e into the iSMC framework is desirable but presents considerable obstacles, and is therefore left as a future perspective.

Our results provide evidence that similarly to humans (Harpak et al., 2016; Smith et al., 2018), the mutation landscape is a crucial determinant of the distribution of diversity in *D. melanogaster*. The simulation study (Figure 5, Figure 7) further suggests that in many evolutionary scenarios the mutation landscape will remain the most relevant factor shaping π along the genome, depending notably on the window size used in the analysis. Future work using integrative models like the ones we introduced here (Figure 1, Figure 3) and applied to species with distinct genomic features and life-history traits will help elucidate how often – and by how much – the mutation landscape stands out as the main driver of nucleotide diversity.

We emphasize that we have not directly argued in favor of either genetic drift or natural selection in the classic population genetics debate, but instead we have highlighted the importance of a third element – the mutation landscape – in shaping genome-wide diversity. Nevertheless, the mutation landscape should play a role in the dynamics of natural selection by modulating the rate at which variation is input into genes (and other functionally important elements) depending on their position in the genome. Consequently, levels of selective interference, genetic load and rates of adaptation should vary accordingly (Castellano et al., 2016). In *D. melanogaster*, our inferred mutation landscape varies ~10-fold between minimum and maximum values at the 50 kb scale, meaning that the impact of mutation rate variation on selective processes can be substantial. This opens intriguing lines of inquiry. For example, under what conditions can the shape of the mutation landscape itself be selected for? It has been shown that modifiers of the global mutation rate are under selection to reduce genetic load (Lynch, 2008; Lynch et al., 2016). It remains to be seen whether the position of genes or genomic features correlated with local μ (e.g., replication timing (Francioli et al., 2015)) can likewise be optimized (Martincorena & Luscombe, 2013).

After all, population genetics theory predicts that at equilibrium the reduction in mean fitness of the population due to recurrent mutations is equal to the sum of mutation rates among sites where they hit deleteriously and actually independent of their selective effects (Haldane, 1937). Curiously, during the editing of this manuscript, the first evidence of an adaptive mutation landscape was reported in *Arabidopsis thaliana*, with coding regions experiencing fewer *de novo* mutations than the rest of the genome, and essential genes even less so (Monroe et al., 2022). This suggests that local mutation rates have been themselves under selective pressure to reduce genetic load in at least one model system, and indicates that perhaps an even smaller fraction of the depletion of nucleotide diversity near genes can be directly attributed to linked selection than previously inferred. In *D. melanogaster*, we failed to find a relationship between local mutation rate and selective constraint (recall that the correlation test between $\hat{\theta}$ and the proportion of exonic sites yielded Spearman's rho = -0.037 with p-value = 0.19, at the 50 kb scale); however, this could also be due to lack of power in the test because of the relatively large window size we used, combined with *Drosophila*'s high gene density. At any rate, much more effort is needed to explore the causes and consequences of mutation rate variation across the tree of life. As a starting point, we can ask how conserved the mutation landscape is in closely-related species (or, equivalently, how fluid is its evolution within populations). Analogous work on the recombination landscape has revealed overall fast evolution of "hotspots" in mammals and has helped uncover the molecular architecture responsible for the placement of double-strand breaks (Berg et al., 2011; Jabbari et al., 2019). Moreover, adaptive dynamics have been evoked to explain the differences in the recombination landscape between populations of *D. pseudoobscura*, (Samuk et al., 2020). It will be interesting to test whether mutation events follow similar patterns, now that the impact of various sequence motifs on local μ is being more thoroughly investigated (DeWitt et al., 2021; Kim et al., 2021; Oman et al., 2022). Unraveling the factors that shape the mutation landscape at different genomic scales will likely provide important insight. For example, can the large-scale variation in mutation rates that we found in *D. melanogaster* be partially explained by aggregation of short (differentially mutable) sequence motifs, or is it driven by independent genomic features? As the molecular underpinnings of adaptive mutation landscapes become elucidated (e.g., what kind of proteins, sequence motifs and epigenetic markers are involved in increasing replication fidelity in functionally constrained regions and eventually decreasing it where polymorphism would tend to be beneficial) we will gain a better understanding of how flexible such phenotype is and how prevalent it is expected to be in different phylogenetic groups. It is plausible that modifiers of the mutation landscape may be successfully optimized, at least in species with high enough N_e for such second-order effects to be seen by selection (Lynch, 2010; Martincorena & Luscombe, 2013; Sung et al., 2012). Recent work notably highlighted the importance of epigenetic factors in shaping the mutation landscape and started to shed light on its evolutionary consequences (Habig et al., 2021; Möller et al., 2021). The variety of molecular agents recruited to tweak the mutation landscape and create pockets of decreased or increased *de novo* mutation rates can be plenty, and it only outlines the complexity of evolutionary biology.

In hindsight, it is perhaps not surprising that mutation rate variation has a profound impact on nucleotide diversity. Mutations are, after all, the "stuff of evolution" (Nei, 2013), and distinct genomic regions displaying differential influx of SNPs must have sharp consequences to the analyses and interpretation of genetic data. This argument is naturally transferred to the ongoing discussion about incorporating complex demography and background selection into the null model of molecular evolution (Comeron, 2017, 2014; Johri et al., 2020), which is motivated by the goal of providing more sensible expectations for rigorously testing alternative scenarios. Our results suggest that a more realistic null model should also include variation in the mutation rate along the genome. By doing so, genome-wide scans (e.g., looking for regions with reduced diversity summary statistics as candidates for selective sweeps) may become less susceptible to both false negatives (in regions of high mutation rate) and false positives (in regions of low mutation rate), paving the way to more robust inference (Booker et al., 2017; Haasl & Payseur, 2016; Venkat et al., 2018).

Acknowledgments

The authors thank Aaron Ragsdale, Chris Kyriazis, Guillaume Achaz, Guy Sella, Kai Zeng, Laurent Duret, Josep Comeron, Kirk Lohmueller and Pier Palamara for discussions about this work; Ana Filipa Moutinho for providing organized data on *Drosophila*. Preprint version 3 has been peer reviewed and recommended by PCI Evolutionary Biology (https://doi.org/10.24072/pci.evolbiol.100636) (Racimo, 2023).

Funding

JYD acknowledges funding from the Max Planck Society. This work was supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft) attributed to JYD, within the priority program (SPP) 1590 “probabilistic structures in evolution”.

Conflicts of Interest Statement

The authors declare they have no financial conflict of interest relating to the content of this article.

Data and Software Availability

The iSMC software package and source code is freely available at https://doi.org/10.5281/zenodo.7826556 (Barroso, 2023a). Scripts used to generate our results can be found at https://doi.org/10.5281/zenodo.7826575 (Barroso, 2023b). Data required to reproduce our results are deposited in FigShare under the DOI https://doi.org/10.6084/m9.figshare.13164320 (Barroso, 2023c). The second repository contains the script ‘dm_analyses.Rmd’. Combined with the data on FigShare, this script generates a PDF document named ‘dm_analyses.pdf’ which contains the results and diagnostics of all linear models and correlations reported in this article.

References

Adrion, J.R., Galloway, J.G., Kern, A.D., 2019. Inferring the landscape of recombination using recurrent neural networks. bioRxiv 662247. https://doi.org/10.1101/662247

Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M., Gerstein, M.B., 2010. Annotating non-coding regions of the genome. Nat. Rev. Genet. 11, 559–571. https://doi.org/10.1038/nrg2814

Andolfatto, P., 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome. Genome Res. 17, 1755–1762. https://doi.org/10.1101/gr.6691007

Baetu, T., 2019. Mechanisms in Molecular Biology. Elem. Philos. Biol. https://doi.org/10.1017/9781108592925

Barroso, G.V., Puzović, N., Dutheil, J.Y., 2019. Inference of recombination maps from a single pair of genomes and its application to ancient samples. PLOS Genet. 15, e1008449. https://doi.org/10.1371/journal.pgen.1008449

Barroso, G.V., 2023a. gvbarroso/iSMC: v0.0.24 (Software package and Source code). Zenodo. https://doi.org/10.5281/zenodo.7826556

Barroso, G.V., 2023b. gvbarroso/ismc_dm_analyses: v0.0.1 (Scripts). Zenodo. https://doi.org/10.5281/zenodo.7826575

Barroso, G.V., 2023c. Quantifying the determinants of the genome-wide diversity in Drosophila using iSMC (Data). FigShare. https://doi.org/10.6084/m9.figshare.13164320

Begun, D.J., Aquadro, C.F., 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature 356, 519–520. https://doi.org/10.1038/356519a0

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.-P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N., Pachter, L., Myers, E., Langley, C.H., 2007. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. PLOS Biol. 5, e310. https://doi.org/10.1371/journal.pbio.0050310

Beichman, A.C., Huerta-Sanchez, E., Lohmueller, K.E., 2018. Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms. Annu. Rev. Ecol. Evol. Syst. 49, 433–456. https://doi.org/10.1146/annurev-ecolsys-110617-062431

Berg, I.L., Neumann, R., Sarbajna, S., Odenthal-Hesse, L., Butler, N.J., Jeffreys, A.J., 2011. Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in

African populations. Proc. Natl. Acad. Sci. U. S. A. 108, 12378–12383. https://doi.org/10.1073/pnas.1109531108

Besenbacher, S., Hvilsom, C., Marques-Bonet, T., Mailund, T., Schierup, M.H., 2019. Direct estimation of mutations in great apes reconciles phylogenetic dating. Nat. Ecol. Evol. 1. https://doi.org/10.1038/s41559-018-0778-x

Booker, T.R., Jackson, B.C., Keightley, P.D., 2017. Detecting positive selection in the genome. BMC Biol. 15, 98. https://doi.org/10.1186/s12915-017-0434-y

Booker, T.R., Yeaman, S., Whitlock, M.C., 2020. Variation in recombination rate affects detection of outliers in genome scans under neutrality. Mol. Ecol. mec.15501. https://doi.org/10.1111/mec.15501

Buffalo, V., 2021. Natural Selection is Unlikely to Explain Why Species Get a Thin Slice of π. bioRxiv 2021.02.03.429633. https://doi.org/10.1101/2021.02.03.429633

Casillas, S., Barbadilla, A., 2017. Molecular Population Genetics. Genetics 205, 1003–1035. https://doi.org/10.1534/genetics.116.196493

Castellano, D., Coronado-Zamora, M., Campos, J.L., Barbadilla, A., Eyre-Walker, A., 2016. Adaptive Evolution Is Substantially Impeded by Hill–Robertson Interference in Drosophila. Mol. Biol. Evol. 33, 442–455. https://doi.org/10.1093/molbev/msv236

Castellano, D., Eyre-Walker, A., Munch, K., 2020. Impact of Mutation Rate and Selection at Linked Sites on DNA Variation across the Genomes of Humans and Other Homininae. Genome Biol. Evol. 12, 3550–3561. https://doi.org/10.1093/gbe/evz215

Castellano, D., James, J., Eyre-Walker, A., 2018. Nearly Neutral Evolution across the Drosophila melanogaster Genome. Mol. Biol. Evol. 35, 2685–2694. https://doi.org/10.1093/molbev/msy164

Chan, A.H., Jenkins, P.A., Song, Y.S., 2012. Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. PLOS Genet. 8, e1003090. https://doi.org/10.1371/journal.pgen.1003090

Charlesworth, B., 2013. Background Selection 20 Years on: The Wilhelmine E. Key 2012 Invitational Lecture. J. Hered. 104, 161–171. https://doi.org/10.1093/jhered/ess136

Charlesworth, B., 2010. Molecular population genomics: a short history. Genet. Res. 92, 397–411. https://doi.org/10.1017/S0016672310000522

Charlesworth, B., 2009. Effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. 10, 195–205. https://doi.org/10.1038/nrg2526

Charlesworth, B., Charlesworth, D., 2017. Population genetics from 1966 to 2016. Heredity 118, 2–9. https://doi.org/10.1038/hdy.2016.55

Charlesworth, B., Morgan, M.T., Charlesworth, D., 1993. The Effect of Deleterious Mutations on Neutral Molecular Variation. Genetics 134, 1289–1303. https://doi.org/10.1093/genetics/134.4.1289

Comeron, J.M., 2017. Background selection as null hypothesis in population genomics: insights and challenges from Drosophila studies. Philos. Trans. R. Soc. B Biol. Sci. 372, 20160471. https://doi.org/10.1098/rstb.2016.0471

Comeron, J.M., 2014. Background Selection as Baseline for Nucleotide Variation across the Drosophila Genome. PLOS Genet. 10, e1004434. https://doi.org/10.1371/journal.pgen.1004434

Comeron, J.M., Ratnappan, R., Bailin, S., 2012. The Many Landscapes of Recombination in Drosophila melanogaster. PLoS Genet. 8, e1002905. https://doi.org/10.1371/journal.pgen.1002905

Cunningham, F., Allen, J.E., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., Donaldson, S., El Houdaigui, B., El Naboulsi, T., Fatima, R., Giron, C.G., Genez, T., Martinez, J.G., Guijarro-Clarke, C., Gymer, A., Hardy, M., Hollis, Z., Hourlier, T., Hunt, T., Juettemann, T., Kaikala, V., Kay, M., Lavidas, I., Le, T., Lemos, D., Marugán, J.C., Mohanan, S., Mushtaq, A., Naven, M., Ogeh, D.N., Parker, A., Parton, A., Perry, M., Piližota, I., Prosovetskaia, I., Sakthivel, M.P., Salam, A.I.A., Schmitt, B.M., Schuilenburg, H., Sheppard, D., Pérez-Silva, J.G., Stark, W., Steed, E., Sutinen, K., Sukumaran, R., Sumathipala, D., Suner, M.-M., Szpak, M., Thormann, A., Tricomi, F.F., Urbina-Gómez, D., Veidenberg, A., Walsh, T.A., Walts, B., Willhoft, N., Winterbottom, A., Wass, E., Chakiachvili, M., Flint, B., Frankish, A., Giorgetti, S., Haggerty, L., Hunt, S.E., IIsley, G.R., Loveland, J.E., Martin, F.J., Moore, B., Mudge, J.M., Muffato, M., Perry, E., Ruffier, M., Tate, J., Thybert, D., Trevanion, S.J., Dyer, S., Harrison, P.W., Howe, K.L., Yates, A.D., Zerbino, D.R., Flicek, P., 2022. Ensembl 2022. Nucleic Acids Res. 50, D988–D995. https://doi.org/10.1093/nar/gkab1049

Cutter, A.D., Payseur, B.A., 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat. Rev. Genet. 14, 262–274. https://doi.org/10.1038/nrg3425

DeWitt, W.S., Harris, K.D., Ragsdale, A.P., Harris, K., 2021. Nonparametric coalescent inference of mutation spectrum history and demography. Proc. Natl. Acad. Sci. 118. https://doi.org/10.1073/pnas.2013798118

Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G., 1998. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511790492

Dutheil, J.Y., 2021. Towards more realistic models of genomes in populations: The Markov-modulated sequentially Markov coalescent, in: Baake, E., Wakolbinger, A. (Eds.), EMS Series of Congress Reports. EMS Press, pp. 383–408. https://doi.org/10.4171/ecr/17-1/18

Dutheil, J.Y., 2017. Hidden Markov Models in Population Genomics. Methods Mol. Biol. Clifton NJ 1552, 149–164. https://doi.org/10.1007/978-1-4939-6753-7_11

Ellegren, H., Galtier, N., 2016. Determinants of genetic diversity. Nat. Rev. Genet. 17, 422–433. https://doi.org/10.1038/nrg.2016.58

Elyashiv, E., Sattath, S., Hu, T.T., Strutsovsky, A., McVicker, G., Andolfatto, P., Coop, G., Sella, G., 2016. A Genomic Map of the Effects of Linked Selection in Drosophila. PLOS Genet. 12, e1006130. https://doi.org/10.1371/journal.pgen.1006130

Felsenstein, J., Churchill, G.A., 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13, 93–104. https://doi.org/10.1093/oxfordjournals.molbev.a025575

Ferré, J., 2009. 3.02 - Regression Diagnostics, in: Brown, S.D., Tauler, R., Walczak, B. (Eds.), Comprehensive Chemometrics. Elsevier, Oxford, pp. 33–89. https://doi.org/10.1016/B978-044452701-1.00076-4

Francioli, L.C., Polak, P.P., Koren, A., Menelaou, A., Chun, S., Renkens, I., Consortium, G. of the N., Duijn, C.M. van, Swertz, M., Wijmenga, C., Ommen, G. van, Slagboom, P.E., Boomsma, D.I., Ye, K., Guryev, V., Arndt, P.F., Kloosterman, W.P., Bakker, P.I.W. de, Sunyaev, S.R., 2015. Genome-wide patterns and properties of de novo mutations in humans. Nat. Genet. 47, 822–826. https://doi.org/10.1038/ng.3292

Galtier, N., Rousselle, M., 2020. How Much Does Ne Vary Among Species? Genetics 216, 559–572. https://doi.org/10.1534/genetics.120.303622

Haasl, R.J., Payseur, B.A., 2016. Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. Mol. Ecol. 25, 5–23. https://doi.org/10.1111/mec.13339

Habig, M., Lorrain, C., Feurtey, A., Komluski, J., Stukenbrock, E.H., 2021. Epigenetic modifications affect the rate of spontaneous mutations in a pathogenic fungus. Nat. Commun. 12, 5869. https://doi.org/10.1038/s41467-021-26108-y

Haldane, J.B.S., 1937. The Effect of Variation of Fitness. Am. Nat. 71, 337–349. https://doi.org/10.1086/280722

Haller, B.C., Messer, P.W., 2018. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. Mol. Biol. Evol. https://doi.org/10.1093/molbev/msy228

Harpak, A., Bhaskar, A., Pritchard, J.K., 2016. Mutation Rate Variation is a Primary Determinant of the Distribution of Allele Frequencies in Humans. PLOS Genet. 12, e1006489. https://doi.org/10.1371/journal.pgen.1006489

Haudry, A., Laurent, S., Kapun, M., 2020. Population Genomics on the Fly: Recent Advances in Drosophila, in: Dutheil, J.Y. (Ed.), Statistical Population Genomics, Methods in Molecular Biology. Springer US, New York, NY, pp. 357–396. https://doi.org/10.1007/978-1-0716-0199-0_15

Hein, J., Schierup, M., Wiuf, C., 2004. Gene Genealogies, Variation and Evolution: A primer in coalescent theory. Oxford University Press, Oxford, New York. https://doi.org/10.1080/10635150500354860

Hey, J., 1999. The neutralist, the fly and the selectionist. Trends Ecol. Evol. 14, 35–38. https://doi.org/10.1016/s0169-5347(98)01497-9

Hodgkinson, A., Eyre-Walker, A., 2011. Variation in the mutation rate across mammalian genomes. Nat. Rev. Genet. 12, 756–766. https://doi.org/10.1038/nrg3098

Hubisz, M.J., Williams, A.L., Siepel, A., 2020. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. PLOS Genet. 16, e1008895. https://doi.org/10.1371/journal.pgen.1008895

Hudson, R.R., 1983. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23, 183–201. https://doi.org/10.1016/0040-5809(83)90013-8

Hudson, R.R., Kaplan, N.L., 1995. Deleterious background selection with recombination. Genetics 141, 1605–1617. https://doi.org/10.1093/genetics/141.4.1605

Hudson, R.R., Kaplan, N.L., 1994. Gene Trees with Background Selection, in: Golding, B. (Ed.), Non-Neutral Evolution: Theories and Molecular Data. Springer US, Boston, MA, pp. 140–153. https://doi.org/10.1007/978-1-4615-2383-3_12

Hudson, R.R., Kaplan, N.L., 1988. The coalescent process in models with selection and recombination. Genetics 120, 831–840. https://doi.org/10.1093/genetics/120.3.831

Hudson, R.R., Kaplan, N.L., 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111, 147–164. https://doi.org/10.1093/genetics/111.1.147

Jabbari, K., Wirtz, J., Rauscher, M., Wiehe, T., 2019. A common genomic code for chromatin architecture and recombination landscape. PLOS ONE 14, e0213278. https://doi.org/10.1371/journal.pone.0213278

Jensen, J.D., Payseur, B.A., Stephan, W., Aquadro, C.F., Lynch, M., Charlesworth, D., Charlesworth, B., 2019. The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. Evolution 73, 111–114. https://doi.org/10.1111/evo.13650

Johri, P., Charlesworth, B., Jensen, J.D., 2020. Toward an Evolutionarily Appropriate Null Model: Jointly Inferring Demography and Purifying Selection. Genetics 215, 173–192. https://doi.org/10.1534/genetics.119.303002

Jónsson, H., Sulem, P., Arnadottir, G.A., Pálsson, G., Eggertsson, H.P., Kristmundsdottir, S., Zink, F., Kehr, B., Hjorleifsson, K.E., Jensson, B.Ö., Jonsdottir, I., Marelsson, S.E., Gudjonsson, S.A., Gylfason, A., Jonasdottir, Adalbjorg, Jonasdottir, Aslaug, Stacey, S.N., Magnusson, O.T., Thorsteinsdottir, U., Masson, G., Kong, A., Halldorsson, B.V., Helgason, A., Gudbjartsson, D.F., Stefansson, K., 2018. Multiple transmissions of de novo mutations in families. Nat. Genet. 50, 1674–1680. https://doi.org/10.1038/s41588-018-0259-9

Kelleher, Jerome, Etheridge, A.M., McVean, Gilean, 2016. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLOS Comput. Biol. 12, e1004842–e1004842. https://doi.org/10.1371/journal.pcbi.1004842

Kelleher, J., Thornton, K.R., Ashander, J., Ralph, P.L., 2018. Efficient pedigree recording for fast population genetics simulation. PLOS Comput. Biol. 14, e1006581. https://doi.org/10.1371/journal.pcbi.1006581

Kern, A.D., Hahn, M.W., 2018. The Neutral Theory in Light of Natural Selection. Mol. Biol. Evol. 35, 1366–1371. https://doi.org/10.1093/molbev/msy092

Kim, Y.-A., Leiserson, M.D.M., Moorjani, P., Sharan, R., Wojtowicz, D., Przytycka, T.M., 2021. Mutational Signatures: From Methods to Mechanisms. Annu. Rev. Biomed. Data Sci. 4, 189–206. https://doi.org/10.1146/annurev-biodatasci-122320-120920

Kimura, M., 1968. Evolutionary Rate at the Molecular Level. Nature 217, 624–626. https://doi.org/10.1038/217624a0

Kingman, J., 1982. On the Genealogy of Large Populations. J. Appl. Probab. 19, 27–43. https://doi.org/10.2307/3213548

Lack, J.B., Cardeno, C.M., Crepeau, M.W., Taylor, W., Corbett-Detig, R.B., Stevens, K.A., Langley, C.H., Pool, J.E., 2015. The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila melanogaster Genomes, Including 197 from a Single Ancestral Range Population. Genetics 199, 1229–1241. https://doi.org/10.1534/genetics.115.174664

Lawrie, D.S., Messer, P.W., Hershberg, R., Petrov, D.A., 2013. Strong Purifying Selection at Synonymous Sites in D. melanogaster. PLOS Genet. 9, e1003527. https://doi.org/10.1371/journal.pgen.1003527

Lewontin, R.C., 1974. The Genetic Basis of Evolutionary Change. Columbia University Press.

Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. Nature 475, 493–496. https://doi.org/10.1038/nature10231

Lynch, M., 2010. Rate, molecular spectrum, and consequences of human mutation. Proc. Natl. Acad. Sci. 107, 961–968. https://doi.org/10.1073/pnas.0912629107

Lynch, M., 2008. The Cellular, Developmental and Population-Genetic Determinants of Mutation-Rate Evolution. Genetics 180, 933–943. https://doi.org/10.1534/genetics.108.090456

Lynch, M., Ackerman, M.S., Gout, J.-F., Long, H., Sung, W., Thomas, W.K., Foster, P.L., 2016. Genetic drift, selection and the evolution of the mutation rate. Nat. Rev. Genet. 17, 704–714. https://doi.org/10.1038/nrg.2016.104

Machado, H.E., Lawrie, D.S., Petrov, D.A., 2020. Pervasive Strong Selection at the Level of Codon Usage Bias in Drosophila melanogaster. Genetics 214, 511–528. https://doi.org/10.1534/genetics.119.302542

Malaspinas, A.-S., Westaway, M.C., Muller, C., Sousa, V.C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J.Y., Crawford, J.E., Heupink, T.H., Macholdt, E., Peischl, S., Rasmussen, S., Schiffels, S., Subramanian, S., Wright4, J.L., Albrechtsen, A., Barbieri, C., Dupanloup, I., Eriksson, A., Margaryan, A., Moltke, I., Pugach, I., Korneliussen, T.S., Levkivsky, I.P., Moreno-Mayar, J.V., Ni, S., Racimo, F., Sikora, M., YaliXue, Aghakhanian, F.A., Brucato, N., SørenBrunak, Campos, P.F., Clark, W., Ellingvåg, S., Fourmile, G., Gerbault, P., Injie, D., Koki, G., Leavesley, M., Logan, B., Lynch, A., Matisoo-Smith, E.A., McAllister, P.J., Mentzer, A.J., Metspalu, M., Migliano, A.B., Murgha, L., Phipps, M.E., Pomat, W., Reynolds, D., Ricaut, F.-X., Siba, P., Thomas, M.G., Wales, T., Wall, C., Oppenheimer, S.J., Tyler-Smith, C., Durbin, R., Dortch, J., Manica, A., Schierup, M.H., Foley, R.A., Lahr, M.M., Bowern, C., Wall, J.D., Mailund, T., Stoneking, M., Nielsen, R., Sandhu, M.S., Excoffier, L., Lambert, D.M., Willerslev, E., 2016. A Genomic History of Aboriginal Australia. Nature 1–20. https://doi.org/10.1038/nature18299

Marjoram, P., Wall, J.D., 2006. Fast "coalescent" simulation. BMC Genet. 7, 16–16. https://doi.org/10.1186/1471-2156-7-16

Martincorena, I., Luscombe, N.M., 2013. Non-random mutation: The evolution of targeted hypermutation and hypomutation. BioEssays 35, 123–130. https://doi.org/10.1002/bies.201200150

Matheson, J., Masel, J., 2023. Unlinked background selection reduces neutral diversity more than linked background selection. bioRxiv. https://doi.org/10.1101/2022.01.11.475913

McGaugh, S.E., Heil, C.S.S., Manzano-Winkler, B., Loewe, L., Goldstein, S., Himmel, T.L., Noor, M.A.F., 2012. Recombination Modulates How Selection Affects Linked Sites in Drosophila. PLOS Biol. 10, e1001422. https://doi.org/10.1371/journal.pbio.1001422

McVean, G., 2007. The structure of linkage disequilibrium around a selective sweep. Genetics 175, 1395–1406. https://doi.org/10.1534/genetics.106.062828

McVean, G.A.T., Cardin, N.J., 2005. Approximating the coalescent with recombination. Philos. Trans. R. Soc. B Biol. Sci. 360, 1387–1393. https://doi.org/10.1098/rstb.2005.1673

McVicker, G., Gordon, D., Davis, C., Green, P., 2009. Widespread Genomic Signatures of Natural Selection in Hominid Evolution. PLOS Genet. 5, e1000471. https://doi.org/10.1371/journal.pgen.1000471

Möller, M., Habig, M., Lorrain, C., Feurtey, A., Haueisen, J., Fagundes, W.C., Alizadeh, A., Freitag, M., Stukenbrock, E.H., 2021. Recent loss of the Dim2 DNA methyltransferase decreases mutation rate in repeats and changes evolutionary trajectory in a fungal pathogen. PLOS Genet. 17, e1009448. https://doi.org/10.1371/journal.pgen.1009448

Monroe, J.G., Srikant, T., Carbonell-Bejerano, P., Becker, C., Lensink, M., Exposito-Alonso, M., Klein, M., Hildebrandt, J., Neumann, M., Kliebenstein, D., Weng, M.-L., Imbert, E., Ågren, J., Rutter, M.T., Fenster, C.B., Weigel, D., 2022. Mutation bias reflects natural selection in Arabidopsis thaliana. Nature 602, 101–105. https://doi.org/10.1038/s41586-021-04269-6

Moutinho, A.F., Trancoso, F.F., Dutheil, J.Y., 2019. The Impact of Protein Architecture on Adaptive Evolution. Mol. Biol. Evol. 36, 2013–2028. https://doi.org/10.1093/molbev/msz134

Murphy, D.A., Elyashiv, E., Amster, G., Sella, G., 2022. Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements. eLife 11, e76065. https://doi.org/10.7554/eLife.76065

Nei, M., 2013. Mutation-Driven Evolution. OUP Oxford.

Nordborg, M., Charlesworth, B., Charlesworth, D., 1996. The effect of recombination on background selection*. Genet. Res. 67, 159–174. https://doi.org/10.1017/S0016672300033619

Ohta, T., 1992. The Nearly Neutral Theory of Molecular Evolution. Annu. Rev. Ecol. Syst. 23, 263–286. https://doi.org/10.1146/annurev.es.23.110192.001403

Oman, M., Alam, A., Ness, R.W., 2022. How Sequence Context-Dependent Mutability Drives Mutation Rate Variation in the Genome. Genome Biol. Evol. 14, evac032. https://doi.org/10.1093/gbe/evac032

Palamara, P.F., Terhorst, J., Song, Y.S., Price, A.L., 2018. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. Nat. Genet. 50, 1311–1317. https://doi.org/10.1038/s41588-018-0177-x

Pearl, J., Mackenzie, D., 2018. The Book of Why: The New Science of Cause and Effect, 1st ed. Basic Books, Inc., USA.

Phung, T.N., Huber, C.D., Lohmueller, K.E., 2016. Determining the Effect of Natural Selection on Linked Neutral Divergence across Species. PLOS Genet. 12, e1006199. https://doi.org/10.1371/journal.pgen.1006199

Pouyet, F., Gilbert, K.J., 2021. Towards an improved understanding of molecular evolution: the relative roles of selection, drift, and everything in between. PEER COMMUNITY Evol. Biol. 22. https://doi.org/10.24072/pcjournal.16

Powell, M.J.D., 1964. An efficient method for finding the minimum of a function of several variables without calculating derivatives. Comput. J. 7, 155–162. https://doi.org/10.1093/comjnl/7.2.155

Racimo, F 2023. An unusual suspect: the mutation landscape as a determinant of local variation in nucleotide diversity (Recommendation). PCI Evolutionary Biology. https://doi.org/10.24072/pci.evolbiol.100636

Rasmussen, M.D., Hubisz, M.J., Gronau, I., Siepel, A., 2014. Genome-Wide Inference of Ancestral Recombination Graphs. PLoS Genet. 10. https://doi.org/10.1371/journal.pgen.1004342

Rosenberg, N.A., Nordborg, M., 2002. Genealogical Trees, Coalescent Theory and the Analysis of Genetic Polymorphisms. Nat. Rev. Genet. 3, 380–390. https://doi.org/10.1038/nrg795

Samuk, K., Manzano-Winkler, B., Ritz, K.R., Noor, M.A.F., 2020. Natural Selection Shapes Variation in Genome-wide Recombination Rate in Drosophila pseudoobscura. Curr. Biol. 30, 1517-1528.e6. https://doi.org/10.1016/j.cub.2020.03.053

Schiffels, S., Durbin, R., 2014. Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46, 919–925. https://doi.org/10.1038/ng.3015

Schiffels, S., Wang, K., 2020. MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent, in: Dutheil, J.Y. (Ed.), Statistical Population Genomics, Methods in Molecular Biology. Springer US, New York, NY, pp. 147–166. https://doi.org/10.1007/978-1-0716-0199-0_7

Schlichta, F., Peischl, S., Excoffier, L., 2022. The Impact of Genetic Surfing on Neutral Genomic Diversity. Mol. Biol. Evol. 39, msac249. https://doi.org/10.1093/molbev/msac249

Sellinger, T.P.P., Awad, D.A., Moest, M., Tellier, A., 2020. Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data. PLOS Genet. 16, e1008698. https://doi.org/10.1371/journal.pgen.1008698

Slatkin, M., 2008. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. 9, 477–485. https://doi.org/10.1038/nrg2361

Smith, J.M., Haigh, J., 1974. The hitch-hiking effect of a favourable gene. Genet. Res. 23, 23–35. https://doi.org/10.1017/S0016672308009579

Smith, T.C.A., Arndt, P.F., Eyre-Walker, A., 2018. Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. PLoS Genet. 14, e1007254. https://doi.org/10.1371/journal.pgen.1007254

Spence, J.P., Steinrücken, M., Terhorst, J., Song, Y.S., 2018. Inference of population history using coalescent HMMs: review and outlook. Curr. Opin. Genet. Dev., Genetics of Human Origins 53, 70–76. https://doi.org/10.1016/j.gde.2018.07.002

Staab, P.R., Zhu, S., Metzler, D., Lunter, G., 2015. Scrm: Efficiently simulating long sequences using the approximated coalescent with recombination. Bioinformatics 31, 1680–1682. https://doi.org/10.1093/bioinformatics/btu861

Stankowski, S., Chase, M.A., Fuiten, A.M., Rodrigues, M.F., Ralph, P.L., Streisfeld, M.A., 2019. Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. PLOS Biol. 17, e3000391. https://doi.org/10.1371/journal.pbio.3000391

Stephan, W., Wiehe, T.H.E., Lenz, M.W., 1992. The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. Theor. Popul. Biol. 41, 237–254. https://doi.org/10.1016/0040-5809(92)90045-U

Stern, A.J., Wilton, P.R., Nielsen, R., 2019. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. PLOS Genet. 15, e1008384. https://doi.org/10.1371/journal.pgen.1008384

Sung, W., Ackerman, M.S., Miller, S.F., Doak, T.G., Lynch, M., 2012. Drift-barrier hypothesis and mutation-rate evolution. Proc. Natl. Acad. Sci. 109, 18488–18492. https://doi.org/10.1073/pnas.1216223109

Terhorst, J., Kamm, J.A., Song, Y.S., 2017. Robust and scalable inference of population history from hundreds of unphased whole-genomes. Nat. Genet. 49, 303–309. https://doi.org/10.1038/ng.3748

Venkat, A., Hahn, M.W., Thornton, J.W., 2018. Multinucleotide mutations cause false inferences of lineage-specific positive selection. Nat. Ecol. Evol. 2, 1280–1288. https://doi.org/10.1038/s41559-018-0584-5

Wiehe, T.H., Stephan, W., 1993. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from Drosophila melanogaster. Mol. Biol. Evol. 10, 842–854. https://doi.org/10.1093/oxfordjournals.molbev.a040046

Wiuf, C., Hein, J., 1999. Recombination as a point process along sequences. Theor. Popul. Biol. 55, 248–59. https://doi.org/10.1006/tpbi.1998.1403

Zeng, K., Charlesworth, B., 2011. The joint effects of background selection and genetic recombination on local gene genealogies. Genetics 189, 251–266. https://doi.org/10.1534/genetics.111.130575

Zeng, K., Jackson, B.C., Barton, H. J. 2018. Methods for estimating demography and detecting between-locus differences in the effective population size and mutation rate. Mol. Biol. Evol. https://doi.org/10.1093/molbev/msy212