TITLE: Mutation rate variation shapes genome-wide diversity in *Drosophila melanogaster*

AUTHORS: Gustavo V. Barroso*[1,2] and Julien Y. Dutheil[1,3]

1) Max Planck Institute for Evolutionary Biology. Department of Evolutionary Genetics. August-Thienemann-Straße 2 24306 Plön – GERMANY

2) Current address: 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

* Corresponding author e-mail: gvbarroso@gmail.com

AUTHOR CONTRIBUTIONS: GVB and JYD conceived and designed the study, conducted the simulations, analysed the data and wrote the manuscript. GVB implemented the method.

ACKNOWLEDGMENTS: The authors thank Chris Kyriazis, Guillaume Achaz, Kai Zeng, Laurent Duret, Josep Comeron, Kirk Lohmueller and Pier Palamara for discussions about this work; Ana Filipa Moutinho for providing organised data on *Drosophila*. 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".

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 formalise a causal model of how these micro-evolutionary forces 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 robust inferences in population genomics.

INTRODUCTION:

Understanding how various evolutionary forces shape nucleotide diversity – typically measured as the average pairwise heterozygosity, π – is a major goal of population genomics (Charlesworth, 2009; Ellegren & Galtier, 2016), with a rich history of theoretical and empirical studies that have the fruit fly *Drosophila melanogaster* at its centre (Charlesworth, 2013; Charlesworth & Charlesworth, 2016; Casillas & Barbadilla, 2017). 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 (Buffalo, 2021; Galtier & Rousselle, 2020). Later on, with recognition that linkage and recombination wrap the genome in regions of correlated evolutionary histories (Hudson, 1983; Kaplan & Hudson, 1985), focus shifted toward understanding how diversity levels vary along chromosomes of a single species (Pouyet & Gilbert, 2020). 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 (Hudson & Kaplan, 1988; Cutter & Payseur, 2013) – at first in terms of selective sweeps (Maynard 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, 1994; Hudson & Kaplan, 1995; 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. Nevertheless, this search has so far been incomplete. The literature has mostly considered selection (Comeron, 2014; Elyashiv et al., 2016; McVicker et al., 2009; Stankowski et al., 2019) and introgression (Schrider et al., 2018; Stankowski et al., 2019; Hubisz et al., 2020) as potential explainers of the variation in π along the genome, whereas spatial variation in *de novo* mutation rates (μ) has been largely ignored, presumably due to challenges in its estimation (Jónsson et al., 2018; Besenbacher et al., 2019). Yet a recent study based on human trios suggests that the impact of the mutation landscape on

polymorphism may be greater than previously recognised: 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 rate at the 100 kb scale (Smith et al., 2018). It is unclear, however, how well these results generalise 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., 2018a; Castellano et al., 2018b). Since these indirect measures of the mutation rate are susceptible to the confounding effect of selection (both direct, *e.g.* codon usage (Lawrie et al., 2013; Machado et al., 2020) and indirect, *e.g.* background selection (Phung et al., 2016)), 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, in a second step, the relative contribution of these micro-evolutionary forces to the distribution of diversity can be more meaningfully quantified.

Disentangling the effects of multiple forces shaping the evolution of DNA sequences is challenging. For example, a genomic region with reduced diversity can be explained by either linked selection, drift, low mutation rate or a combination thereof. Zeng and Jackson developed a likelihood-based framework that jointly infers the effective population size (Ne) (Charlesworth, 2009) and μ with high accuracy (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 (Slatkin, 2008), it discards sites in the genome where local variation in the mutation rate may be relevant. In this article we describe 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 and variation in the recombination rate along the genome (Barroso et al., 2019) via a Markov-modulated

Markov process. We now extend this framework to 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). Because natural selection disturbs τ away from its distribution under neutrality around functionally constrained regions of the genome (Palamara et al., 2018; Rasmussen et al., 2014; Stern et al., 2019), iSMC offers estimators of all micro-evolutionary forces and we can further use causal inference (Pearl & Mackenzie, 2018) to simultaneously estimate their effects on π. Our analyses of *Drosophila melanogaster* reveals the impact of linked selection, however, it suggests that the rate of *de novo* mutations is quantitatively the most important factor shaping nucleotide diversity in this species.

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 (McVean & Cardin, 2005; Marjoram & Wall, 2006). Its first implementation as an inference tool derives the transition probabilities of genealogies between adjacent sites as a function of the historical variation in Ne and the genome-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, 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, where the local ρ is 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 a HMM where the hidden states are *n*-tuples storing all combinations of genealogies and discretised 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 optimised, *e.g.* by maximum likelihood (see Methods). In the iSMC with heterogeneous recombination (ρ-iSMC) the hidden states are 2-tuples containing 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 algorithm of the HMM, respectively. Parameter optimisation and subsequent posterior decoding is performed as detailed by (Barroso et al., 2019). Under strict neutrality (Ne homogeneous along the genome), 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.

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; it does not, however, explicitly enforce that they have common genomic landscapes. Rather, iSMC uses posterior probabilities to reconstruct recombination and mutation maps separately for each diploid.

The large variance of the coalescent results in variation among the posterior landscapes inferred from each diploid because their specific τ landscapes modulate the expected number of SNPs they contain in different regions of the genome (hence the amount of information they bear, in those regions, about ancestral processes such as mutation and recombination). To reduce noise from the individual 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 (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, noting, 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.

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

We sought to quantify the determinants of genome-wide diversity in *Drosophila 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. Owing to the complexity of the model,

which led to large computation times, 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 estimated parameters suggested an exponential-like distribution of recombination rates ($\hat{\alpha} = \hat{\beta} = 1.03$ for their Gamma distribution) whereas the inferred distribution of mutation rates was more tightly centered around the mean ($\hat{\alpha} = \hat{\beta} = 2.93$ for their Gamma distribution). iSMC also inferred that the change in recombination rate across the genome was more frequent (auto-correlation parameter $\hat{\delta}_\rho \sim 0.9999$, corresponding to a change of recombination rate on average every 10 kb) than the change in mutation rate (auto-correlation parameter $\hat{\delta}_\theta \sim 0.99999$, corresponding to a change of mutation rate on average every 100 kb). This suggests that our model is mostly sensitive to factors that determine large-scale variation in the mutation rate instead of fine-scale sequence motifs such as highly mutable triplets (DeWitt et al., 2021; Harris & Pritchard, 2017). Our estimated genome-wide average $\hat{\rho}$ (0.036) is in line with previous estimates (Chan et al., 2012), and the coalescence rates (which, for the purpose of this paper, comprise a collection of nuisance parameters used to refine our estimates of τ, ρ and θ along the genome) suggest a relatively recent ~4-fold bottleneck followed by a partial recovery (**Supplemental Figure S1**). We used these parameters estimated from *D. melanogaster* to simulate 10 replicate datasets under a purely neutral scenario (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 analyse the determinants of nucleotide variation at different scales by binning the landscapes 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 rank correlation, hereafter referred to as Spearman's rho, **Figure 2A,**

Supplemental Table S1), showing 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 forces shape the distribution of nucleotide diversity along the genome, measured as π, the average 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 may thus cast our causal model as an ordinary least squares regression (OLS) that seeks to explain π as a linear combination of the centered variables ρ, θ and τ and statistical associations between our explanatory variables and π then represent causal relationships that merit scientific explanation. Since simulations grant direct access to the true genomic landscapes, the ensuing OLS models are free of estimation noise in the explanatory variables and serve as a ground truth assessment of how neutral evolutionary forces linearly influence nucleotide diversity. In all replicates, we found that model selection using Akaike's information criterion favours a regression with an interaction term between the two variables that directly influence π $\pi_i = \beta_1 \cdot \tau_i + \beta_2 \cdot \theta_i + \beta_3 \cdot \rho_i + \beta_4 \cdot \theta : \tau + \epsilon_i$ over the simpler model $\pi_i = \beta_1 \cdot \tau_i + \beta_2 \cdot \theta_i + \beta_3 \cdot \rho_i + \epsilon_i$.

Fitting this 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. This is expected since both deeper ancestry and higher mutation rate lead to increased diversity and the effect 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, 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). 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 predicted by our causal model (**Figure 3**). To quantify the relative influence of drift, mutation and recombination to local diversity levels, we used type II ANOVA to partition the R^2 contributed by each explanatory variable. Our estimates show that the θ landscape explains most of the variance in π in our simulatios 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 very fine 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) makes it comparatively more relevant at larger windows. To test whether we could recover such trends with our HMM, we fitted the OLS models to the same landscapes of nucleotide diversity except using the ρ, θ and τ landscapes inferred by iSMC (instead of the true, simulated ones). The sign and significance of the estimated coefficients of the OLS models remained unchanged, 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 our estimates of the linear coefficients, we also fitted Generalized Least Squares (GLS) models accounting for both effects, which reassuringly produced consistent results. Although co-linearity between $\hat{\tau}$ and $\hat{\theta}$ 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). Type II ANOVA using the inferred landscapes shows that the contribution of $\hat{\tau}$ is slightly higher than using the true landscapes (5.1%, 2.9% and 1.4%, increasing window size) whereas the contribution of 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**). We conclude that the joint-inference approach of iSMC can infer the genomic landscapes of τ, ρ and θ

and that the OLS representation of our causal model (**Figure 3**) can quantify their effect on the distribution of nucleotide diversity π.

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 causal model assumes that sequencing errors are unbiased with respect to the explanatory variables and that the population is broadly panmictic. We used our inferred maps to fit an OLS regression of the form $\pi_i = \beta_1 \cdot \hat{\tau}_i + \beta_2 \cdot \hat{\theta}_i + \beta_3 \cdot \hat{\rho}_i + \beta_4 \cdot \hat{\theta} : \hat{\tau} + \epsilon_i$. As in our simulations, the regression model shows positive effects of both τ and θ, but not of ρ, on π across all scales (**Table 1**). Likewise, a GLS model yields the same trends, and the variance inflation factors are < 5, indicating that the estimated coefficients are robust to co-linearity (Ferré, 2009). 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 effect 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 in our neutral simulations, however, in real *Drosophila* data the simple correlation test between $\hat{\rho}$ and π ends up positive and significant, 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 and Aquadro (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 in SMC-based methods τ

is primarily influenced by demography (fluctuating population sizes explicitly by use of a Coalescent prior taming the transition probabilities of the HMM (Li & Durbin, 2011; Schiffels & Durbin, 2014) and population structure implicitly, because lineages may only coalesce within the same deme, see *e.g.* (Beichman et al., 2018)), it has also been demonstrated to carry the signature of selection (Palamara et al., 2018) due to local changes in coalescence rates that have been interpreted as spatial variation in Ne (Charlesworth, 2009). We tested the sensitivity of our regression framework to this effect by fitting linear models without $\hat{\tau}$ as an explanatory variable $\pi_i = \beta_1 \cdot \hat{\theta}_i + \beta_2 \cdot \hat{\rho}_i + \epsilon_i$, 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**, lower panel), 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 selection on the diversity landscape is smaller than that of mutation rate variation.

To investigate the signature of selection, we analysed the relationship between the local mutation rate and the levels of synonymous (π_S) and non-synonymous (π_N) diversity across *D. melanogaster* genes (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 their correlation tests cannot be explained by a bias in $\hat{\theta}$ 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 a bias in our inference of the mutation rate in regions under stronger background selection (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 = 2.03e-12, partial correlation accounting for $\hat{\theta}$), as expected since background selection should reduce the TMRCA more abruptly in densely constrained regions (Palamara et al., 2018). To estimate the effect of these factors on more heavily constrained diversity, we 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 $\hat{\tau}$ and $\hat{\theta}$, 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.

Mutation rate variation shapes genome-wide diversity in several neutral scenarios

Our analyses of *Drosophila* data and *Drosophila*-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 first addressed this question by exploring the parameter space of our neutral simulations. For fixed values of the long-term average population size (Ne = 100,000), the average mutation rate per site per generation (μ = 2e-09), the Gamma distribution of scaling factors of θ (α = β = 2.5) and the Gamma distribution of scaling factors of ρ (α = β = 1.0), we varied the demographic history (flat Ne; 10-fold bottleneck 0.5 coalescent time units ago), the average recombination rate per site per generation (*r* = 1e-08; 1e-09) and the average length of genomic blocks of constant μ (50 kb; 500 kb; flat mutation landscape). We reasoned that

the extent of the variation in τ along the genome compared to that of μ 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 Ne varies in time, the distribution of coalescence times may become 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 it being more homogeneous along the genome, for 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 our 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 (supporting the causal effect

$\rho \rightarrow \tau$ 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).

More generally, our simulation study of neutral scenarios shows that the relative impacts of evolutionary forces 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 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 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 (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, 2014) 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 before, we fitted a ρ-θ-iSMC model with five mutation rate classes, five recombination rate classes and 30 coalescence time intervals, and 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

rank correlations 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).

We report strong and positive correlations between simulated and inferred landscapes under such complex scenario (**Figure 6**). As expected, the inherently fine-scale variation in the distribution of genealogies is the hardest to reconstruct: the Spearman correlation between the true TMRCA landscape and the inferred one ranges from 0.385 to 0.465 at 50 kb and increase substantially with window size (up to 0.787 at 1 Mb, **Supplementary Table S5**). In comparison, the correlation between the true and inferred mutation landscapes ranges from 0.751 (at 1 Mb) to 0.894 (at 50 kb, **Supplementary Table S6**). The recombination landscape is also well recovered, with correlation coefficients ranging from 0.830 (at 50 kb) to 0.963 (at 1Mb, **Supplementary Table S7**). 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 increases correlation among genealogies around constrained sites, 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 (*c.f.* **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 ANOVA 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 OLS models. Moreover, there is a closer agreement of real *Drosophila* data with neutral simulations than with simulations of background selection. A tentative explanation is unrealistically strong linked selection in the

simulations, mismatching the extent of linked selection in real data by a number of possible causes. This would 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 simulations under neutrality (**Figure 4A**, **Figure 5**). We inquired into this phenomenon visually, by looking into the landscapes simulated under background selection, which provide critical insight into the interplay among micro-evolutionary forces shaping diversity (**Figure 7**). 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. 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. Taken together, the results from simulations under background selection 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 simulation results clearly demonstrate that the effect of linked selection can be overwhelmed by mutation rate variation (**Figure 7**). We conclude that the modelling 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 recognised 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 from other species 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. Such 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, Jensen et al., 2019; Kern & Hahn, 2018), with the influence of local mutation rate only recently brought to light (Castellano, Eyre-Walker, 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 *Drosophila* 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.

These results do not imply that linked selection cannot extend beyond the 18.3% of the *Drosophila* 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 *Drosophila* can be explained 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 at least some of these effects. We believe that the discrepancy between our results and Cameron's can be mostly attributed to two important differences between our approaches. First, the B-value statistic computed in (Comeron, 2014) does not account for heterogeneity in the

mutation rate along the genome, unlike its original application (McVicker et al., 2009). Second, the 70% figure derives from the correlation between π and B-value maps alone, without accounting for other causal factors. The B-value represents the expected reduction of diversity solely due to selection against linked deleterious mutations. In our model, this effect is captured by the local TMRCA. Interestingly, we find that simple correlations between π and $\hat{\tau}$ are of the same order (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, *c.f.* Table 1 in (Comeron, 2014)). The central point being that our ANOVA was able to identify θ as the main driver of π only because its effect was jointly estimated with those of τ and ρ in the linear models. Taking a step back, it is also conceivable that selection is not only manifested as distortions in the distribution of genealogies (represented here by the $\hat{\tau}$ landscape) but also biases our estimate of the mutation landscape. However, based on the high similarity between real *Drosophila* data and our neutral simulations (**Figure 4A**) as well as on iSMC estimates' robustness to the presence of linked selection (**Figure 2B**, **Figure 4B**, **Figure 6**), we argue that such bias is unlikely 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 and Kaplan, 1994), which we have not explored because estimation noise with iSMC is typically higher at such small window sizes. Shortly, our results provide evidence that similarly to humans (Harpak et al., 2016; Smith et al., 2018), the mutation landscape is a crucial driver of the distribution of diversity in *D. melanogater*. Our 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 studies 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.

Our results do not directly argue in favor of either genetic drift or natural selection in the classic population genetics debate, but instead highlight 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) according to 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 optimised. 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 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. 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 *Drosophila 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 and sequence motifs 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 optimised, at least in species with high enough Ne for such second-order effects to be seen by selection (Lynch, 2010; 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 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, 2014, 2017; 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).

METHODS:

Modelling spatial variation in θ

We now introduce our approach to modelling 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 sequence where 0 represents homozygous states and 1 represents heterozygous states (thus, once genome pairs are established, phasing information is discarded). The probability of observing 0 or 1 at any given position of the genome depends only on τ and the scaled mutation rate θ. If the hidden state configuration of the model excludes 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 compute for every site as $P(0|\tau)=e^{(-\theta\times\tau)}$, and $P(1|\tau)=1-e^{(-\theta\times\tau)}$ respectively, as presented by Li & Durbin (2011).

To incorporate spatial heterogeneity in the mutation rate, we set the genome-wide average θ as the average number of pair-wise differences between haplotypes, and modulate it by drawing scaling factors from a discretised Gamma distribution with mean equal to 1. The parameters shaping this prior distribution are estimated by maximum likelihood together with other parameters of the model. 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 discretised τ intervals, and $n^{(\theta)}$ be the number of discretised 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:

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

where δ_θ is the 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 at genomic position *i* can be written as:

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

where θ_m is the product of the genome-wide average mutation rate and the value of the m-th discretised 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 $P(1|\tau_t, \theta_m) = 1 - e^{(-\theta_m \times \tau_t)}$. The forward recursion integrates over all $n^{(\theta)}$ categories of θ and over all $n^{(\tau)}$ intervals of τ. In the double-modulated model, 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). Since spatial variation in ρ contributes to the transition probability between genealogies, the forward recursion is now given by:

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

(2)

To obtain the single-nucleotide landscapes for a given diploid individual, we first compute the posterior probability of each hidden state for every site *i* in the genome using standard HMM procedures (Durbin et al., 1998). Afterwards, since in ρ-θ-iSMC the hidden states are triplets (**Figure 1**), computing the posterior average of each variable of interest – ρ, θ or τ – amounts to first marginalising the probability distribution of its categories and then using it to weight the corresponding discretised values (Barroso et al., 2019). Let $\hat{M}$ be the inferred discretised 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 the value of $\hat{M}$ inside category *l*. Similarly, let $\hat{R}$ be the inferred discretised 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 the value of $\hat{R}$ inside category *k*. Then the posterior average $\hat{\theta}$ at position *i* is given by:

$$\hat{\bar{\theta}}_i = \hat{\theta}_0 \cdot \sum_{l=1}^{n^{(\theta)}} m_l \cdot \left(\sum_{k=1}^{n^{(\rho)}} \sum_{j=1}^{n^{(\tau)}} P_i(\theta_l, \rho_k, \tau_j) \right) \quad \textbf{(3)}$$

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

$$\hat{\bar{\rho}}_i = \hat{\rho}_0 \cdot \sum_{k=1}^{n^{(\rho)}} r_k \cdot \left(\sum_{l=1}^{n^{(\theta)}} \sum_{j=1}^{n^{(\tau)}} P_i(\theta_l, \rho_k, \tau_j) \right) \quad \textbf{(4)}$$

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

$$\hat{\bar{\tau}}_i = \sum_{j=1}^{n^{(\tau)}} \hat{\tau}_j \cdot \left(\sum_{k=1}^{n^{(\rho)}} \sum_{l=1}^{n^{(\theta)}} P_i(m_l, r_k, \tau_j) \right) \quad \textbf{(5)}$$

We can then bin the inferred single-nucleotide landscapes into 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.*:

$$\hat{\bar{\theta}}^L_{pop} = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{1}{L} \sum_{i=1}^{L} \hat{\theta}_{i,j} \right) \quad \textbf{(6)}$$

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 analysed by iSMC, and likewise for ρ and τ:

$$\hat{\rho}^{L}_{pop} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{L}\sum_{i=1}^{L}\hat{\rho}_{i,j}\right) \quad \textbf{(7)}$$

$$\hat{\tau}^{L}_{pop} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{L}\sum_{i=1}^{L}\hat{\tau}_{i,j}\right) \quad \textbf{(8)}$$

Simulation study

Using SCRM (Staab et al., 2015), we simulated 10 haploid sequences of length 30 Mb with parameters identical to those inferred from ρ-θ-iSMC in *D. melanogaster*: θ = 0.0112; ρ = 0.036; α (continuous Gamma distribution used as mutation rate prior) = 3.0; α (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. **Figures 4** and **5** display the mean R^2 value over 10 replicates, but the standard deviation of these estimates are very small and confidence intervals were therefore omitted.

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 use Comeron's recombination map in 100 kb windows (Comeron, 2014). 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, the mutation and

recombination rates were scaled 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 randomly 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 of mean 1.0 and with a shape parameter equal to 2.5, in segments with lengths drawn from a geometric distribution with mean 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 *Drosophila melanogaster* data was 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). Gaps and unknown nucleotides in these FASTA sequences were assigned as missing data. For each scale in which the landscapes were binned (50 kb, 200 kb and 1 Mb), we filtered out windows with more than 10% missing data in the resulting maps. 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. Genomic coordinates for coding sequences and their summary statistics (π_N, and π_S) were taken from (Moutinho et al., 2019).

Linear modelling

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 S2, S3, S4**). Since the trends in coefficients are overall consistent, we pulled the autosome arms and in the Results section we presented 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 centered all explanatory variables before fitting the regression models to aid in both computation of variance inflation factors and interpretation of the coefficients (due to the interaction term between θ and τ).

DATA AND SOFTWARE AVAILABILITY:

The iSMC software package and source code is freely available at https://github.com/gvbarroso/iSMC

Scripts used to generate our results can be found in https://github.com/gvbarroso/ismc_dm_analyses

Data required to reproduce our results are deposited in FigShare under the DOI 10.6084/m9.figshare.13164320

CONFLICTS OF INTEREST STATEMENT:

The authors declare no competing financial interests.

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 model structure. OLS = Ordinary Least Squares; GLS = Generalised Least Squares.

	Variable	50 kb		200 kb		1 Mb	
		Coefficient	**p-value**	**Coefficient**	**p-value**	**Coefficient**	**p-value**
OLS	θ	0.9746	<2.2e-16	0.9767	<2.2e-16	0.9903	<2.2e-16
	τ	0.0114	<2.2e-16	0.0151	<2.2e-16	0.0098	<2.2e-16
	ρ	0.0018	0.0237	0.0016	0.2580	0.0049	0.0748
	θ:τ	1.0510	<2.2e-16	0.9481	<2.2e-16	0.5321	<1e-4
GLS	θ	0.9730	<1e-4	0.9729	<1e-4	0.9816	<1e-4
	τ	0.0115	<1e-4	0.0114	<1e-4	0.0106	<1e-4
	ρ	0.0014	0.0700	0.0018	0.2300	0.0005	0.8658
	θ:τ	1.0777	<1e-4	0.9323	<1e-4	0.6656	<1e-4
GLS	θ	1.0991	<1e-4	1.0796	<1e-4	1.0840	<1e-4
	ρ	0.0529	<1e-4	0.0601	<1e-4	0.0444	<1e-4

FIGURE LEGENDS:

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). Discretised values of these distributions (left) are combined in triplets as the hidden states of our Hidden Markov Model (right).

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 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 directly 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 model systems where selection affects μ and *r* through modifier genes are shown by dashed lines.

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.

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. In **A** and **B**, results are displayed according simulated parameters (rows = global 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.

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

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 $4 \times Ne$ 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**). These data were extracted straight from the simulations and are, therefore, free of estimation noise.

REFERENCES:

Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M., & Gerstein, M. B. (2010). Annotating non-coding regions of the genome. *Nature Reviews Genetics*, *11*(8), Article 8. https://doi.org/10.1038/nrg2814

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 Genetics*, *15*(11), e1008449. https://doi.org/10.1371/journal.pgen.1008449

Begun, D. J., & Aquadro, C. F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature*, *356*(6369), 519–520. https://doi.org/10.1038/356519a0

Beichman, A. C., Huerta-Sanchez, E., & Lohmueller, K. E. (2018). Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms. *Annual Review of Ecology, Evolution, and Systematics*, *49*(1), 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. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(30), 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. *Nature Ecology & Evolution*, 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 Biology*, *15*(1), 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. *Molecular Ecology*, mec.15501. https://doi.org/10.1111/mec.15501

Buffalo, V. (2021). Why do species get a thin slice of π? Revisiting Lewontin's Paradox of Variation. *BioRxiv*, 2021.02.03.429633. https://doi.org/10.1101/2021.02.03.429633

Casillas, S., & Barbadilla, A. (2017). Molecular Population Genetics. *Genetics*, *205*(3), 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. *Molecular Biology and Evolution*, *33*(2), 442–455. https://doi.org/10.1093/molbev/msv236

Castellano, D., Eyre-Walker, A., & Munch, K. (2018). Impact of mutation rate and selection at linked sites on fine-scale DNA variation across the homininae genome. *BioRxiv*, 452201. https://doi.org/10.1101/452201

Castellano, D., James, J., & Eyre-Walker, A. (2018). Nearly neutral evolution across the Drosophila melanogaster genome. *Molecular Biology and Evolution*, *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 Genetics*, *8*(12), e1003090. https://doi.org/10.1371/journal.pgen.1003090

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*(3), 195–205. https://doi.org/10.1038/nrg2526

Charlesworth, B., & Charlesworth, D. (2016). Population genetics from 1966 to 2016. *Heredity*, *February*, 1–8. https://doi.org/10.1038/hdy.2016.55

Comeron, J. M. (2014). Background Selection as Baseline for Nucleotide Variation across the Drosophila Genome. *PLOS Genetics*, *10*(6), e1004434. https://doi.org/10.1371/journal.pgen.1004434

Comeron, J. M. (2017). Background selection as null hypothesis in population genomics: Insights and challenges from Drosophila studies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1736), 20160471. https://doi.org/10.1098/rstb.2016.0471

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., … Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Research*, *50*(D1), 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. *Nature Reviews. Genetics*, *14*(4), 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. *Proceedings of the National Academy of Sciences*, *118*(21). 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. https://doi.org/10.1017/CBO9780511790492

Dutheil, J. Y. (2017). Hidden Markov Models in Population Genomics. *Methods in Molecular Biology (Clifton, N.J.)*, *1552*, 149–164. https://doi.org/10.1007/978-1-4939-6753-7_11

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

Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews. Genetics*, *17*(7), 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 Genetics*, *12*(8), 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. *Molecular Biology and Evolution*, *13*(1), 93–104. https://doi.org/10.1093/oxfordjournals.molbev.a025575

Ferré, J. (2009). Regression Diagnostics. In *Comprehensive Chemometrics* (Vol. 3, pp. 33–89). Elsevier. https://www.sciencedirect.com/topics/mathematics/variance-inflation-factor

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. *Nature Genetics*, *47*(7), 822–826. https://doi.org/10.1038/ng.3292

Galtier, N., & Rousselle, M. (2020). How Much Does Ne Vary Among Species? *Genetics*, *216*(2), 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. *Molecular Ecology*, *25*(1), 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. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-26108-y

Haldane, J. B. S. (1937). The Effect of Variation of Fitness. *The American Naturalist*, *71*(735), 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. *Molecular Biology and Evolution*. 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 Genetics*, *12*(12), e1006489. https://doi.org/10.1371/journal.pgen.1006489

Harris, K., & Pritchard, J. K. (2017). Rapid evolution of the human mutation spectrum. *ELife*, *6*, e24284. https://doi.org/10.7554/eLife.24284

Hein, J., Schierup, M., & Wiuf, C. (2004). *Gene Genealogies, Variation and Evolution: A primer in coalescent theory*. Oxford University Press.

Hey, J. (1999). The neutralist, the fly and the selectionist. *Trends in Ecology & Evolution*, *14*(1), 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. *Nature Reviews. Genetics*, *12*(11), 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 Genetics*, *16*(8), e1008895. https://doi.org/10.1371/journal.pgen.1008895

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, *23*(2), 183–201. https://doi.org/10.1016/0040-5809(83)90013-8

Hudson, R. R., & Kaplan, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics*, *120*(3), 831–840.

Jabbari, K., Wirtz, J., Rauscher, M., & Wiehe, T. (2019). A common genomic code for chromatin architecture and recombination landscape. *PLOS ONE*, *14*(3), 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*(1), 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*(1), 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, A., Jonasdottir, A., Stacey, S. N., Magnusson, O. T., Thorsteinsdottir, U., … Stefansson, K. (2018). Multiple transmissions of de novo mutations in families. *Nature Genetics*, *50*(12), Article 12. https://doi.org/10.1038/s41588-018-0259-9

Kaplan, N., & Hudson, R. R. (1985). The use of sample genealogies for studying a selectively neutral m-loci model with recombination. *Theoretical Population Biology*, *28*(3), 382–396. https://doi.org/10.1016/0040-5809(85)90036-X

Kelleher, J., Etheridge, A. M., McVean, G., Kingman, J., Hudson, R., Wakeley, J., Hudson, R., Hudson, R., Hudson, R., Kaplan, N., Kaplan, N., Hudson, R., Griffiths, R., Marjoram, P., Donnelly, P., Tavaré, S., Ethier, S., Griffiths, R., Hudson, R., … O'Fallon, B. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*, *12*(5), 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 Computational Biology*, *14*(11), e1006581. https://doi.org/10.1371/journal.pcbi.1006581

Kern, A. D., & Hahn, M. W. (2018). The Neutral Theory in Light of Natural Selection. *Molecular Biology and Evolution*, *35*(6), 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. *Annual Review of Biomedical Data Science*, *4*(1), 189–206. https://doi.org/10.1146/annurev-biodatasci-122320-120920

Kimura, M. (1968). Evolutionary Rate at the Molecular Level. *Nature*, *217*(5129), 624–626. https://doi.org/10.1038/217624a0

Kingman, J. (1982). Genealogy Populations. *Journal of Applied Probability*, *19*(1982), 27–43.

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*(4), 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 Genetics*, *9*(5), 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*(7357), 493–496. https://doi.org/10.1038/nature10231

Lynch, M. (2008). The Cellular, Developmental and Population-Genetic Determinants of Mutation-Rate Evolution. *Genetics*, *180*(2), 933–943. https://doi.org/10.1534/genetics.108.090456

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics: TIG*, *26*(8), 345–352. https://doi.org/10.1016/j.tig.2010.05.003

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. *Nature Reviews Genetics*, *17*(11), Article 11. 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*(2), 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., … 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 Genetics*, *7*(1), 16–16. https://doi.org/10.1186/1471-2156-7-16

McVean, G. A. T., & Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1459), 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 Genetics*, *5*(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 Genetics*, *17*(3), 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*(7895), Article 7895. 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. *Molecular Biology and Evolution*, *36*(9), 2013–2028. https://doi.org/10.1093/molbev/msz134

Nei, M. (2013). *Mutation-Driven Evolution*. OUP Oxford.

Ohta, T. (1992). The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics*, *23*(1), 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 Biology and Evolution*, *14*(3), 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. *Nature Genetics*, *50*(9), 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.

Phung, T. N., Huber, C. D., & Lohmueller, K. E. (2016). Determining the Effect of Natural Selection on Linked Neutral Divergence across Species. *PLOS Genetics*, *12*(8), e1006199. https://doi.org/10.1371/journal.pgen.1006199

Pouyet, F., & Gilbert, K. J. (2020). Towards an improved understanding of molecular evolution: The relative roles of selection, drift, and everything in between. *PEER COMMUNITY IN EVOLUTIONARY BIOLOGY*, 22.

Rasmussen, M. D., Hubisz, M. J., Gronau, I., & Siepel, A. (2014). Genome-Wide Inference of Ancestral Recombination Graphs. *PLoS Genetics*, *10*(5). https://doi.org/10.1371/journal.pgen.1004342

Rosenberg, N. A., & Nordborg, M. (2002). Genealogical Trees, Coalescent Theory and the Analysis of Genetic Polymorphisms. *Nature Reviews Genetics*, *3*(5), 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. *Current Biology*, *30*(8), 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. *Nature Genetics*, *46*(8), 919–925. https://doi.org/10.1038/ng.3015

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

Schrider, D. R., Ayroles, J., Matute, D. R., & Kern, A. D. (2018). Supervised machine learning reveals introgressed loci in the genomes of Drosophila simulans and D. sechellia. *PLOS Genetics*, *14*(4), e1007341. https://doi.org/10.1371/journal.pgen.1007341

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 Genetics*, *16*(4), e1008698. https://doi.org/10.1371/journal.pgen.1008698

Slatkin, M. (2008). Linkage disequilibrium—Understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, *9*(6), 477–485. https://doi.org/10.1038/nrg2361

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 Genetics*, *14*(3), 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. *Current Opinion in Genetics & Development*, *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*(10), 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 Biology*, *17*(7), e3000391. https://doi.org/10.1371/journal.pbio.3000391

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 Genetics*, *15*(9), 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. *Proceedings of the National Academy of Sciences*, *109*(45), 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. *Nature Genetics*, *49*(2), 303–309. https://doi.org/10.1038/ng.3748

Zeng, K., & Jackson, B. C. (2018). Methods for estimating demography and detecting between-locus differences in the effective population size and mutation rate. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msy212

FIGURES:

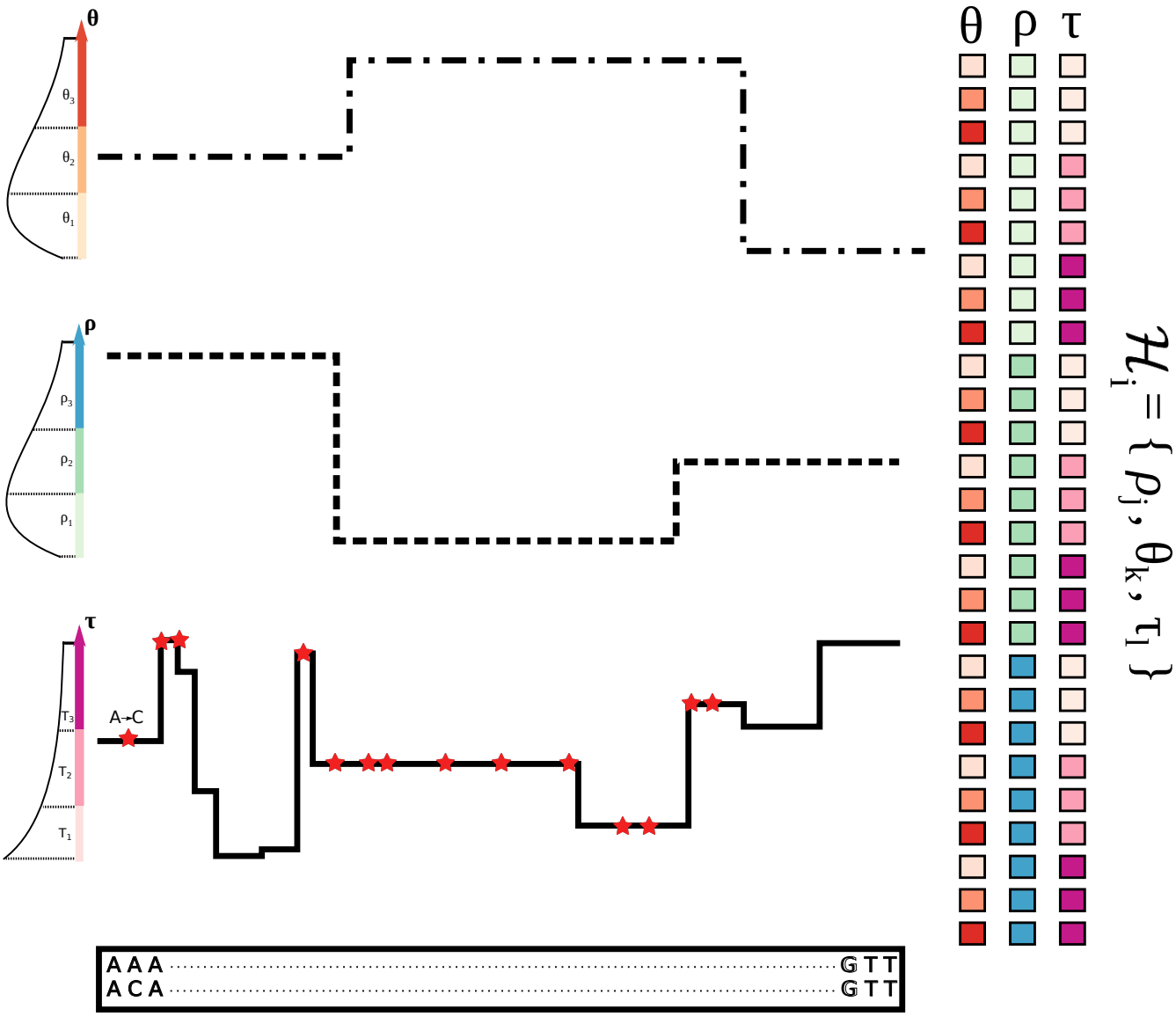

Figure 1. Schematic representation of ρ-θ-iSMC for one pair of genomes. This cartoon model has three time intervals, three recombination rate categories and three mutation rate categories. The genome-wide distribution of diversity depends on the mutation landscape (top) and on the τ landscape (bottom),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

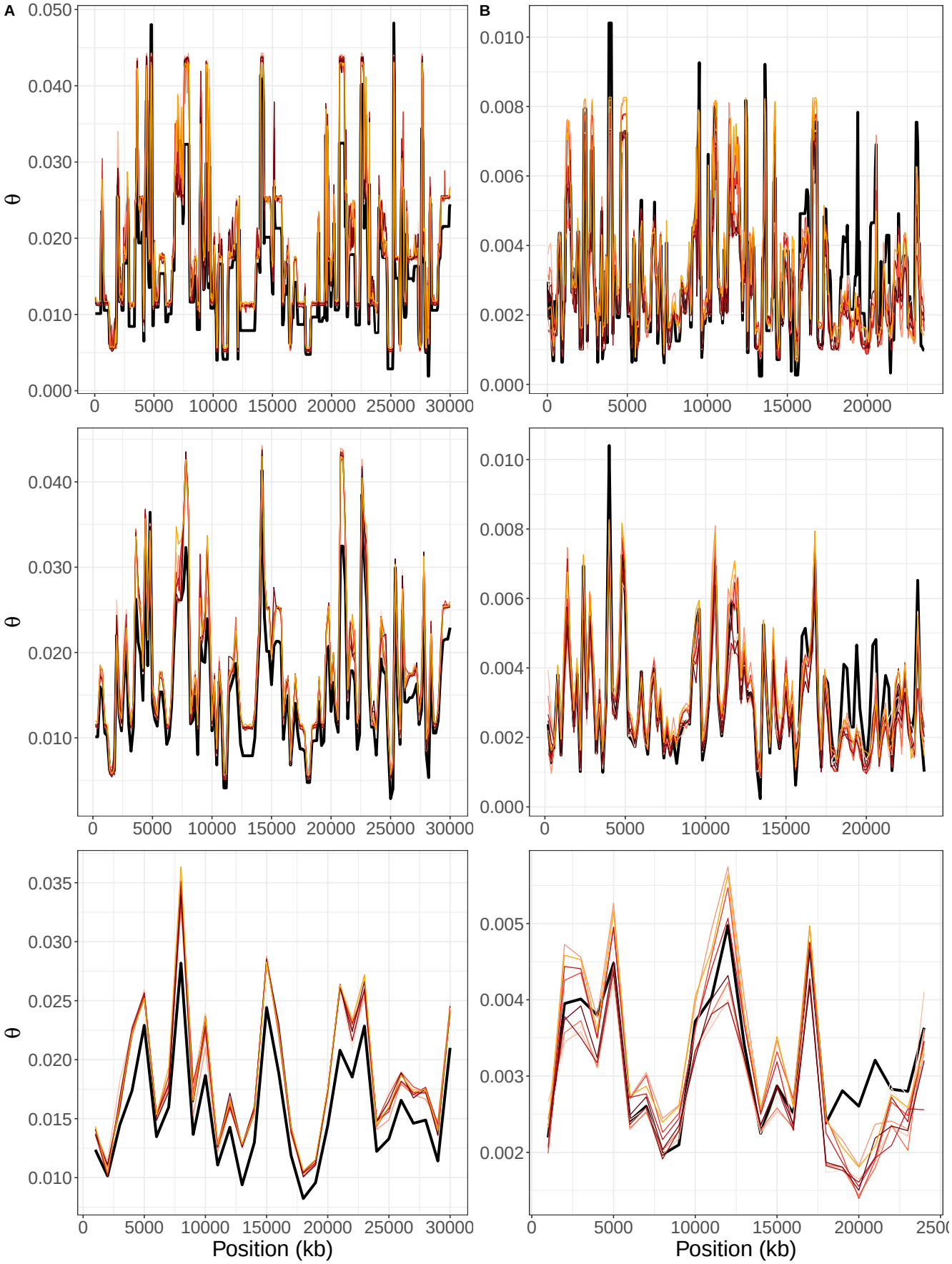
A
B
θ
0.050
0.040
0.030
0.020
0.010
0.000
0.008
0.006
0.004
0.002
0.035
0.025
0.015
0.005
0
5000
10000
15000
20000
25000
30000
Position (kb)

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.

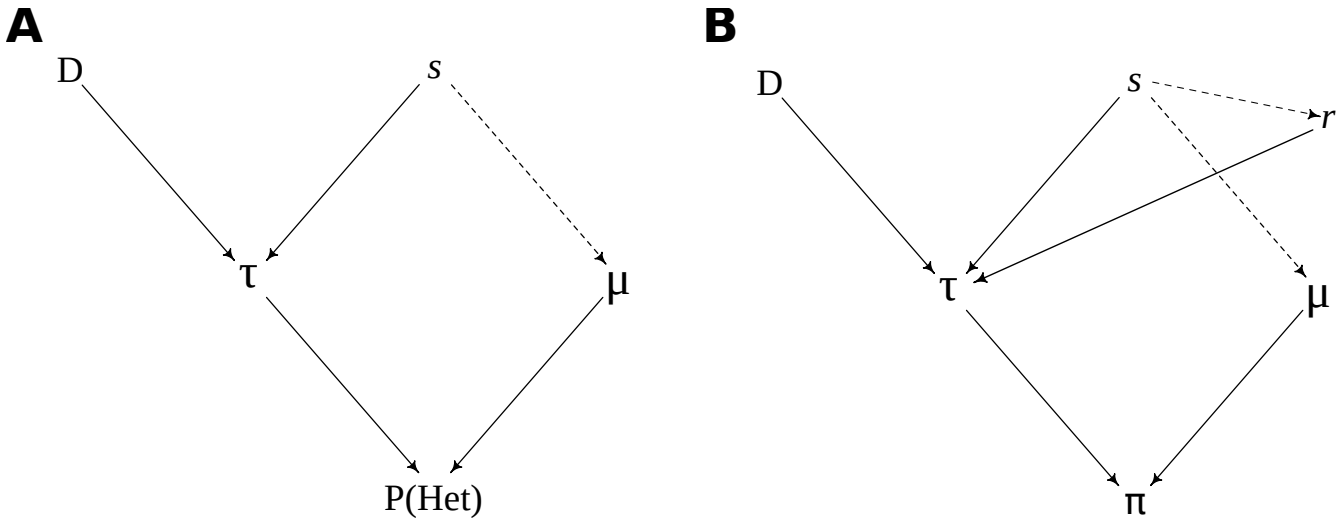

Figure 3. Directed acyclic graphs depicting our causal model for the determinants of genome-wide diversity. A) for a single, hypothetical nucleotide that is independent of any neighbors, its probability of being heterozygous is directly influenced by the local mutation rate (μ) and TMRCA (τ), which in 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 model systems where selection affects μ and *r* through modifier genes are shown by dashed lines.

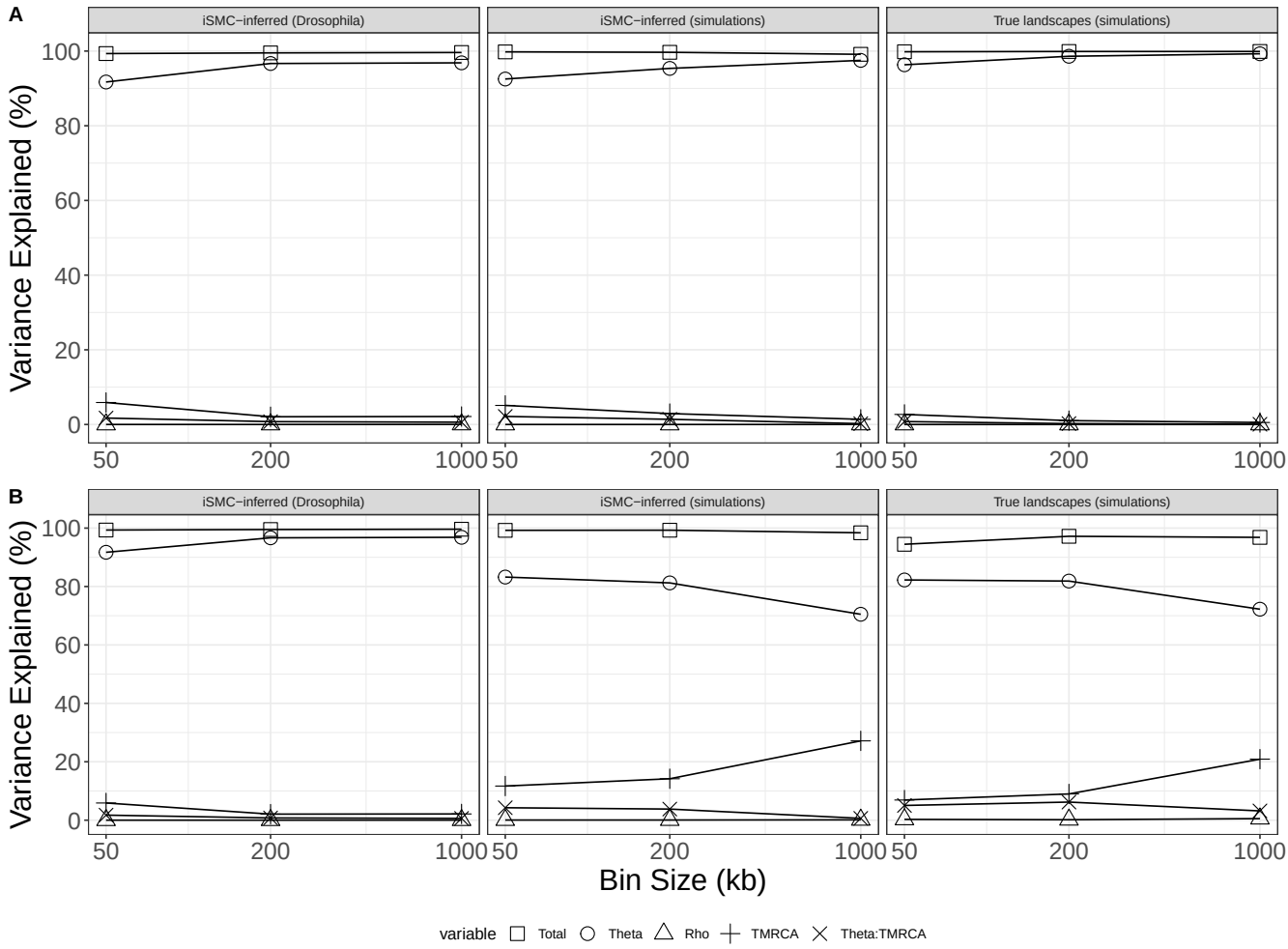

Figure 4. Variance in the distribution of diversity explained by each genomic landscape. Partitioning of variance according to window size (x-axis, shown in $\log_{10}$ scale), using either simulated data (*true* landscapes: right panels; *inferred* landscapes: middle panels) or real *Drosophila* data (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.

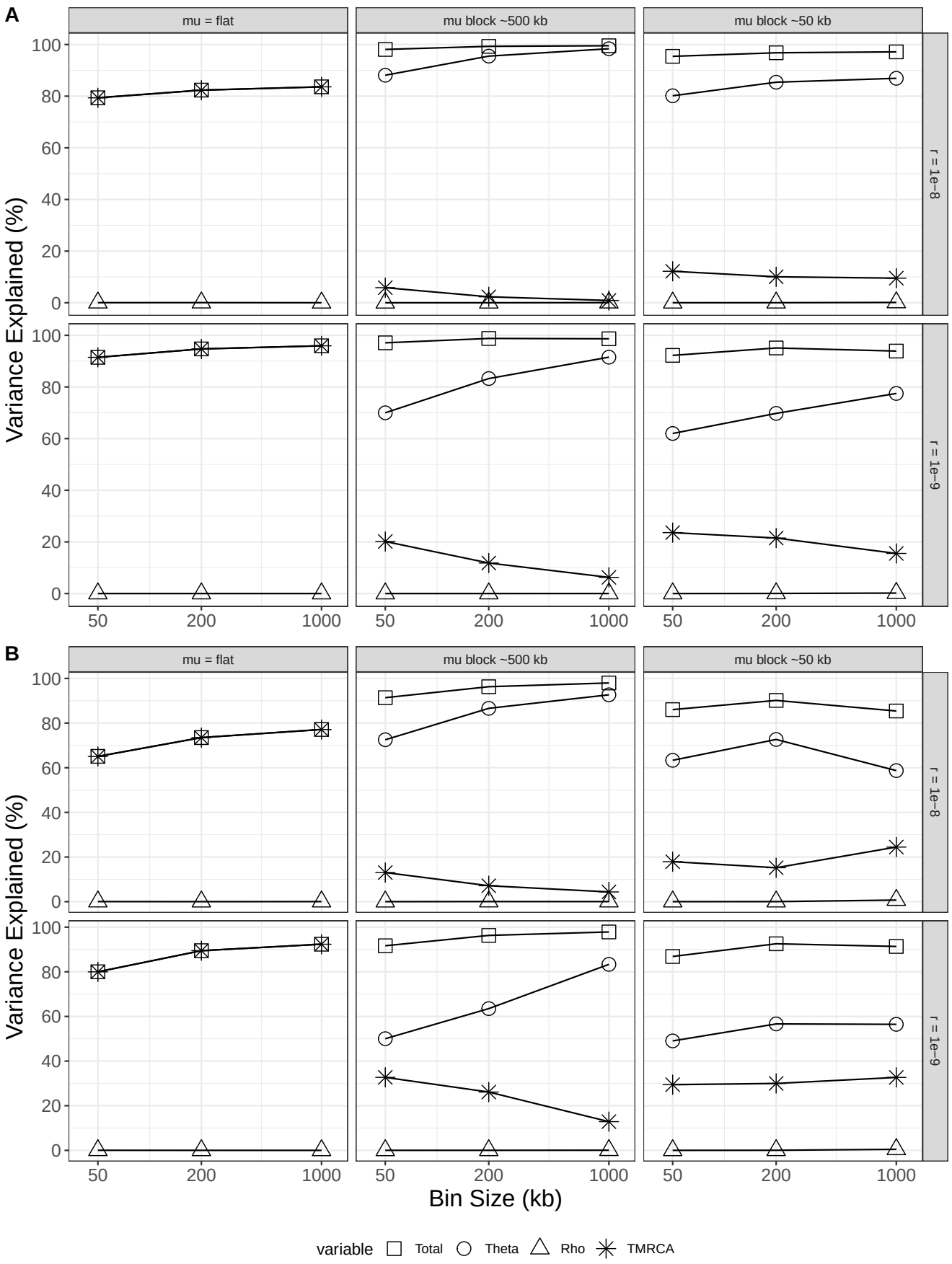
A
mu = flat
mu block ~500 kb
mu block ~50 kb
r = 1e−8
r = 1e−9
Variance Explained (%)
0
20
40
60
80
100
50
200
1000
B
Bin Size (kb)
variable
Total
Theta
Rho
TMRCA

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. In **A** and **B**, results are displayed according simulated parameters (rows = global 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.

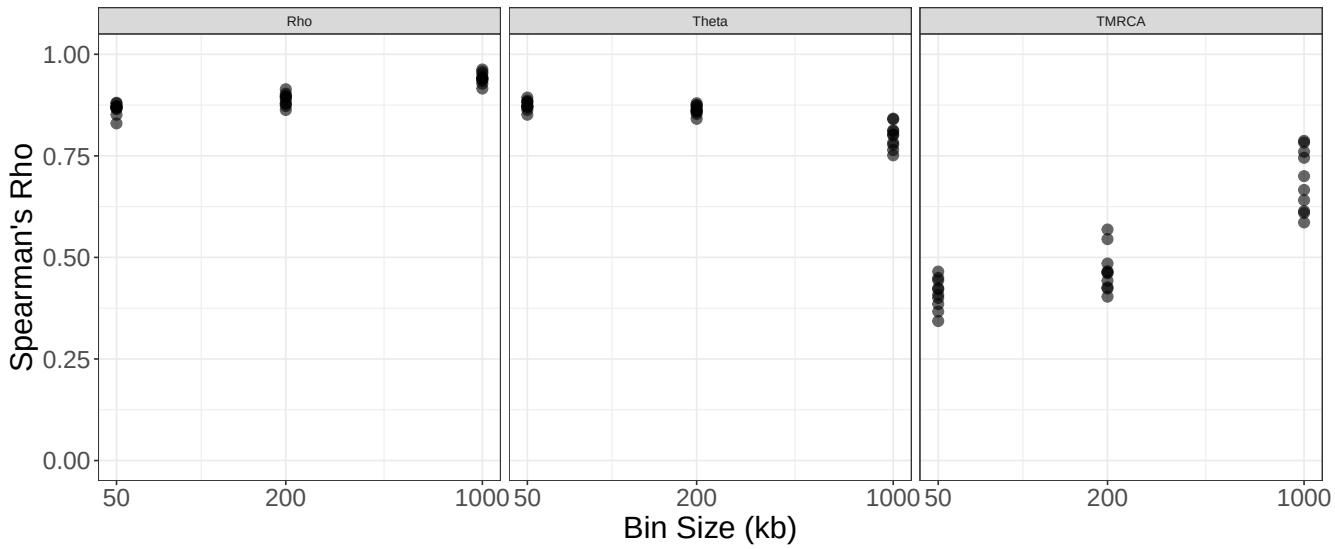

Figure 6. Spearman correlations between inferred and simulated landscapes under background selection. All p-values are smaller than 1e-4.

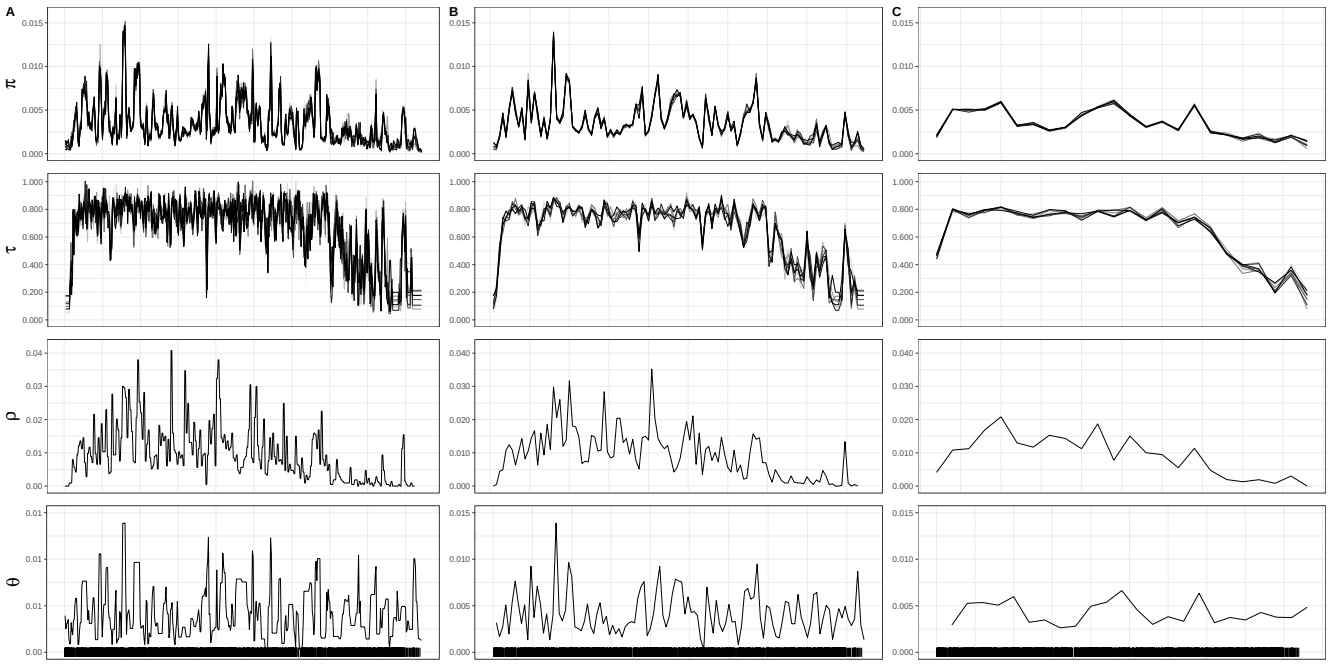

Figure 7. Genomic landscapes simulated under background selection and their effect on the distribution of diversity. From top to bottom: observed nucleotide diversity (lines in shades of grey represent replicates), average TMRCA of the genealogies in units of $4 \times Ne$ generations (lines in shades 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**). These data were extracted straight from the simulations and are, therefore, free of estimation noise.