G3, 2022, jkac236
https://doi.org/10.1093/g3journal/jkac236
Advance Access Publication Date: 14 September 2022
Investigation

Gene flow biases population genetic inference of recombination rate

Kieran Samuk [1,2,*] Mohamed A.F. Noor [1]

[1]Department of Biology, Duke University, Durham, NC 27708, USA,
[2]Department of Evolution, Ecology, and Organismal Biology, The University of California, Riverside,Riverside, CA 92521, USA

*Corresponding author: Department of Evolution, Ecology, and Organismal Biology, The University of California, Riverside, Riverside, CA 92521, USA.
Email: ksamuk@ucr.edu

Abstract

Accurate estimates of the rate of recombination are key to understanding a host of evolutionary processes as well as the evolution of the recombination rate itself. Model-based population genetic methods that infer recombination rates from patterns of linkage disequilibrium in the genome have become a popular method to estimate rates of recombination. However, these linkage disequilibrium-based methods make a variety of simplifying assumptions about the populations of interest that are often not met in natural populations. One such assumption is the absence of gene flow from other populations. Here, we use forward-time population genetic simulations of isolation-with-migration scenarios to explore how gene flow affects the accuracy of linkage disequilibrium-based estimators of recombination rate. We find that moderate levels of gene flow can result in either the overestimation or underestimation of recombination rates by up to 20–50% depending on the timing of divergence. We also find that these biases can affect the detection of interpopulation differences in recombination rate, causing both false positives and false negatives depending on the scenario. We discuss future possibilities for mitigating these biases and recommend that investigators exercise caution and confirm that their study populations meet assumptions before deploying these methods.

Keywords: recombination; population genetics; gene flow; linkage disequilibrium; methods

Introduction

Recombination rate, the number of crossovers per unit genome per generation, plays a key role in shaping evolutionary processes and diversity in the genome. For example, through the action of linked selection, local rates of recombination are a chief determinant of patterns of genetic diversity throughout the genome (Begun and Aquadro 1992; Haddrill *et al.* 2014; Burri 2017; Cutter 2019; Korunes *et al.* 2021). Genome-wide rates of recombination also modulate diverse processes such as adaptation, speciation, and introgression (Samuk *et al.* 2017; Dapper and Payseur 2017; Stapley *et al.* 2017; Schumer *et al.* 2018). There is also a growing appreciation that recombination rate is itself a trait that varies and evolves (Dumont and Payseur 2008; Hunter *et al.* 2016; Johnston *et al.* 2016; Ritz *et al.* 2017; Stapley *et al.* 2017; Samuk *et al.* 2020). Accordingly, there has been great interest in efficient and accurate methods for estimating recombination rates.

Current methods for estimating recombination rates fall into 2 broad classes of methods: direct and indirect (Peñalba and Wolf 2020). Of the direct measures, the 3 most popular approaches are linkage mapping, gamete sequencing, and cytological methods. With classical linkage mapping, map distances between genetic markers are measured by quantifying recombinant markers in the context of a genetic cross or pedigree (Broman 2010; Rastas 2017). The resolution of this approach is limited only by marker density and the sample size of individuals, but larger sample sizes can be grueling to achieve in the laboratory or unavailable in some populations. Furthermore, identifying suitable diagnostic mapping markers can be limiting in some cases (e.g. in a highly homozygous population; Broman 2010). Direct sequencing of pools of recombinant gamete genomes from single individuals using long/linked read sequencing is a newer approach that alleviates many of the issues of traditional mapping, but still requires differentiated markers to score crossover events between homologous chromosomes (Dréau *et al.* 2019; Rommel Fuentes *et al.* 2020; Xu *et al.* 2020). Cytological methods bypass this requirement by directly visualizing recombination-associated protein complexes in cell populations undergoing meiosis (Peterson *et al.* 2019; Peterson and Payseur 2021). However, the cytological methods are limited by the spatial resolution at which such visualization can occur (e.g. the resolution of immunostained gamete karyotypes; Peterson *et al.* 2019).

Because all direct methods of measuring recombination rates are fairly laborious, there has been increased interest in indirect measures of recombination rate that leverage readily available population genetic data. Chief among these are model-based methods that infer rates of recombination from patterns of linkage disequilibrium (LD; Auton and McVean 2007; Chan *et al.* 2012; Kamm *et al.* 2016; Spence and Song 2019). These methods attempt to estimate recombination rates by statistically fitting recombination rates (derived from population genetic models/simulations) to observed

Received: September 27, 2021. **Accepted:** August 30, 2022

patterns of LD. Rather than inferring recombination rate directly, LD-based estimators infer a *population scaled recombination rate*, $\rho = 4N_e r$, where N_e is the effective population size and r is the theoretical per-generation recombination rate. LD-based methods are attractive because they (1) generally only require population-scale genomic data and (2) are very fast, often only requiring several computational hours or less (Spence and Song 2019) and (3) are informative of time-averaged population historical recombination rates (McVean and Auton 2007). Accordingly, LD-based estimates of recombination rates have become extremely popular, and now vastly outnumber direct measures in the literature (Stapley *et al.* 2017; Peñalba and Wolf 2020). These methods have also begun to be used to perform interpopulation comparisons of recombination rates (Stapley *et al.* 2017; Peñalba and Wolf 2020).

Like all models, LD-based estimators of recombination rate make a variety of simplifying assumptions about the populations of interest. For one, they generally assume that the populations/loci of interest are evolving largely neutrally and have reached population genetic equilibrium in a number of ways (Stumpf and McVean 2003). In particular, most methods assume that the populations being studied have reached an equilibrium between recombination and population scaled mutation, such that LD accurately reflects patterns of recombination rate (McVean 2007). Furthermore, it is generally assumed that any form of selection that might distort patterns of LD (e.g. sweeps) has not recently occurred (Chan *et al.* 2012). Finally, these methods make the general assumption that demographic processes that distort genome-wide patterns of LD, such as population size changes, have not occurred (recall that ρ is directly dependent on N_e; Auton and McVean 2007).

While some of these assumptions may be robust to violation, work has shown that some violations can result in biased estimates. For example, Dapper and Payseur (2018) showed that recombination estimates from LDhat (McVean and Auton 2007) are highly sensitive to changes in population size. This can be ameliorated in some cases by incorporating known changes in population size into the estimation procedure, such as implemented in the software pyrho (Spence and Song 2019).

Along with changes in population size and selection, another process that can greatly alter patterns of LD is gene flow. Gene flow and subsequent admixture between diverged populations can have complex effects on patterns of LD within each population (Nei and Li 1973; Ohta 1982). These effects range from large and genomically variable increases in LD due to segregation of divergent haplotypes, to genome-wide decreases in LD as populations become coupled and increase local N_e (Nei and Li 1973; Ohta 1982). While it is now widely accepted that gene flow is commonplace in natural populations (Barton 2001; Mallet 2005; Waples and Gaggiotti 2006; Suvorov *et al.* 2022), there has not been a systematic study of the effects of gene flow on LD-based measures of recombination. Furthermore, it remains unclear how gene flow (or any other violation of assumptions) impacts our ability to detect differences in recombination rate *between* (as opposed to within) populations using LD-based methods.

Here, we address these issues using forward-time population genetic simulations. We attempt to answer 2 specific questions. First, how does gene flow between populations affect the precision and accuracy of LD-based estimates of recombination rate within populations? Secondly, how does gene flow affect our ability to detect evolved differences in recombination rate between populations? Our primary goal is to answer these questions in the context of a core set of realistic demographic scenarios, and not perform an exhaustive exploration of parameter space. Overall, we hope to help investigators understand key sources of bias in LD-based estimates of recombination rate in natural populations and highlight areas of future development.

Methods

Code availability

All scripts used in the analyses described below are available as a repository on Github (http://github.com/ksamuk/LD_recomb).

Forward time simulations with SLiM

To explore how the timing and amount of gene flow affect estimates of recombination rate, we performed forward-time population genetic simulations using SLiM version 3.3 (Haller and Messer 2019). The basic form of all the simulations was an isolation-with-migration scenario: a single ancestral population diverges into 2 subpopulations with a static amount of bidirectional gene flow (Fig. 1). Populations were composed of diploid individuals with 100-kb genomes arranged in a single chromosome. We used genome-wide average estimates of effective population size, mutation rate, and empirical recombination rate from natural populations of *Drosophila melanogaster* (Adrion, Cole *et al.* 2020): Per site mutation rate $= 5.49 \times 10^{-9}$ (Li and Stephan 2006); per site recombination rate $= 2.23 \times 10^{-8}$, (average of chromosome 2R; Comeron *et al.* 2012); N_e = 1.72M (Li and Stephan 2006). Recombination and mutation rates were conservatively modeled as uniform across the 100-kb genome. Following standard practice for forward-time simulations, all simulations were run with an in silico population size of $N = 1{,}000$, and simulated mutation and recombination rates scaled by a factor of N_e/N as per the SLiM manual (Haller and Messer 2019). Note that generation times are also subject to scaling, and for simplicity, we will refer to all generations in terms of back-transformed actual generations rather than SLiM generations (1 SLiM generation ≈ 1,751 actual generations with our scaling factor).

Parameter space

To explore how variation in gene flow affects estimates of recombination, we varied the amount of gene flow over 5 orders of magnitude: 0, 0.01, 0.1, 1, 10, 100, in standard units of N_em (the product of the effective population size and the migration rate). These values were chosen to encompass total isolation ($N_em = 0$), limited gene flow ($N_em = 0.01$–0.1), moderate gene flow in interconnected metapopulations ($N_em = 1$–10; Morjan and Rieseberg 2004; Waples and Gaggiotti 2006), and a scenario of a nascent hybrid swarm ($N_em = 100$). We also varied the timing of the onset of gene flow, with gene flow beginning either immediately after divergence or after a period of isolation. We performed preliminary simulations to determine a period of isolation (~1.7M generations in our case) that produced levels of genomic divergence (Supplementary Fig. 1) similar to those observed in natural population pairs that exhibit genome-wide genetic divergence but still actively exchange genes ($F_{ST} \sim 0.4$; Morjan and Rieseberg 2004; Roux *et al.* 2016). Finally, to explore how gene flow impacts the detection of population *differences* in recombination rate, we modeled scenarios where recombination rate either remains constant in both subpopulations or instantaneously increases by a factor of 2 at the time of divergence in one of the 2 subpopulations (always subpopulation 2). This magnitude of this difference is well within the range of variation in recombination rate reported for a wide variety of species (Stapley *et al.* 2017). In biological terms, an instantaneous increase in population recombination rate could be readily mediated by an environmental change (e.g. temperature, Lloyd *et al.* 2018), a change in mating system (Brandvain and Wright 2016), or whole-genome duplication (Tiley and Burleigh 2015). We note that this instantaneous change

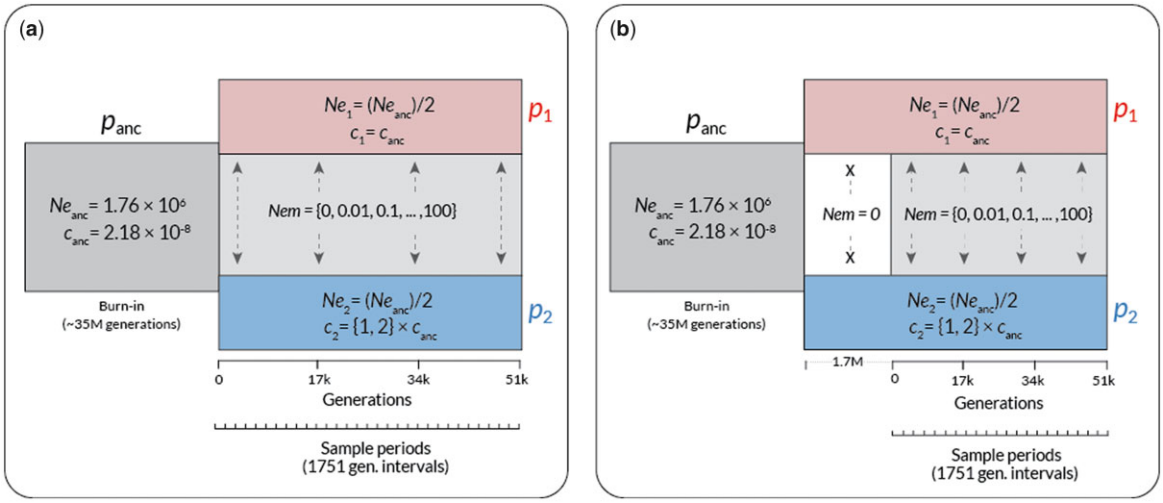

Fig. 1. The structure of the forward-time simulations performed in SLiM. Time in back-transformed generations is shown along the x-axis, and the populations in existence at a given time are shown as rectangles. p_{anc} = the ancestral population, p_1 = the subpopulation with unchanged recombination rate, and p_2 = the subpopulation with increased recombination rate (if applicable). Effective population sizes (N_e) and recombination rates (c, in units of cM/Mb) are shown for each population, with the values for the subpopulations shown relative to the ancestral value. Variable elements of the simulation are shown in braces. Time in generations postdivergence is indicated below the plots, with the precontact isolation period in (b) shown as a dotted line preceding the main axis. Sample periods indicate intervals at which genotypes were output for analysis.

is a "best case" scenario for detecting interpopulation differences in recombination rate, and thus any loss of power to detect differences in recombination that occurs due to gene flow will be conservative.

Details of demographic events

Each simulation began with a single population of size Ne_{anc}, which evolved for a 35M generation burn-in period (following the general practice of a 10-20 N_e burn-in period; Haller and Messer 2019). This initial period was followed by divergence into 2 subpopulations, each with size $Ne_{anc}/2$. Gene flow (for cases where $N_em > 0$) began immediately at the time of divergence or after a 1.7M generation period of isolation and was symmetrical in magnitude and bidirectional. Changes in recombination rate occurred at the time of divergence and instantaneously applied to all individuals in subpopulation 2 only.

Starting at the time of divergence and thereafter in intervals of 1,751 generations, we collected a random sample of 25 individuals (a total of 50 haploid genomes) from each population and saved their complete genotypic at all sites in VCF format. We stopped the simulations after 51,000 generations. Each parameter combination was replicated 100 times, for a total number of $\sim n = 48{,}000$ population samples.

Estimation of recombination rate using pyrho

While there are a variety of LD-based estimators of recombination rate, we elected to use pyrho (Spence and Song 2019) for estimation in this study. It shares its statistical foundation with the most widely used LD-based estimators (LD-hat and LD-helmet; McVean and Auton 2007; Chan *et al.* 2012) while also having the ability to account for changes in effective population size such as we are modeling here (Spence and Song 2019). As such, any estimation biases caused by gene flow will likely affect those approaches as least as much as they affect pyrho. Direct comparisons with other methods are complicated by the fact that pyrho is the only model-based method that adequately accounts for changes in effective population (Adrion, Galloway *et al.* 2020).

We followed the recommended practices for inferring recombination rate using pyrho (https://github.com/popgenmethods/pyrho). We parameterized the initial lookup tables using the effective population size and mutation rates used in the simulations (*unscaled* in this case). To account for changes in effective population size, we created lookup tables that accounted for a change of $N_e/2$ (1.72M to 8.6M) in time steps of 1,751 generations in the past. This allowed us to have an appropriately timed lookup table for each step of the simulation. We used the built-in methods to infer the hyperparameters of window size (best fit 100) and block penalty (best fit 1,000). Using this baseline, we inferred recombination rates using the genotype data (VCF format) from both subpopulations at each time point, for a total of ~96,000 pyrho fits. All computations were performed using the Duke University Computing Cluster, running CentOS Version 8.

Statistical analyses

We performed all data processing and visualization using the tools of the tidyverse package in R 4.0.3 (Wickham 2017; R Core Team 2018). To examine how gene flow between populations affects the accuracy of LD-based estimates of recombination, and the context of the various factors explored in our simulations, we performed an analysis of variance using a linear mixed model with Gaussian errors fitted via the lmer() function from the lme4 package (Bates *et al.* 2015). This model had the following form: Recombination rate = (1|simulation replicate) + (1|simulation generation) + gene flow magnitude + recombination rate change, where (1|[factor]) denotes a random intercept. All variables were standardized (mean-centered and scaled by standard deviation) prior to analysis. To simplify interpretation, we fitted separate models for the continuous gene flow and secondary contact scenarios.

Results

Inference when recombination rate is identical between populations

When the recombination rate remained constant between diverging populations, we found that gene flow introduced 2 types of systematic biases in estimates of recombination rate within populations (Fig. 2a). These effects began when $N_em \geq 1$ in both the

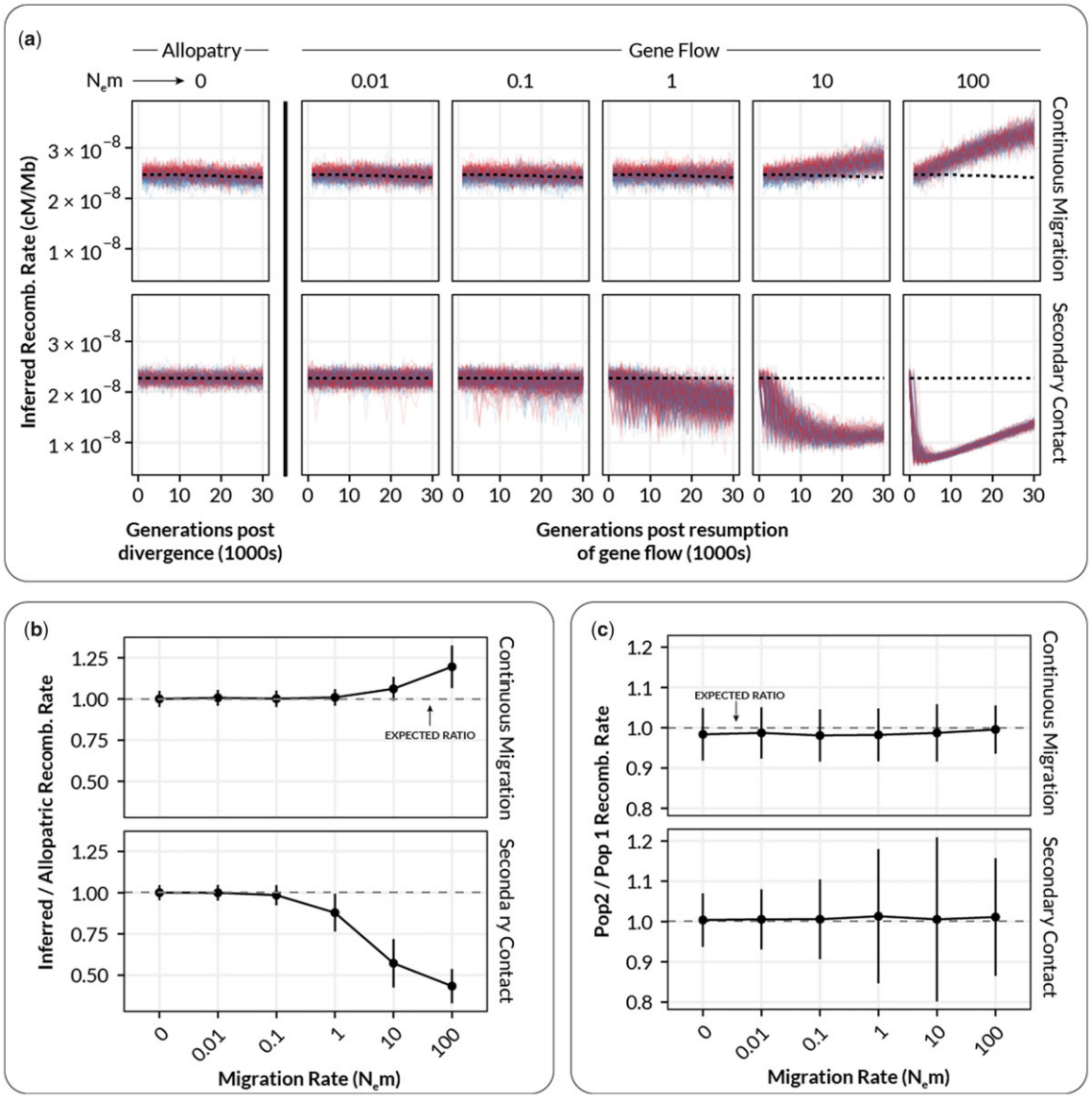

Fig. 2. The relationship between inferred recombination rate and the migration rate in simulated populations where recombination rate remains constant in both subpopulations. (a) Inferred recombination rates for individual simulations at varying levels of migration. Each plot shows inferred rates for simulation replicates (transparent lines) of population 1 (red, unchanged recombination) and population 2 (blue, increased recombination) for a single migration rate. Dashed lines show the expected inferred value in the absence of gene flow (inferred from $N_em = 0$). (b) Summarized inferred recombination rates (*y*-axis) for each level of migration (*x*-axis) from the simulations in a. Points are mean values and error bars depict standard deviations (summarized across all generations). Dashed lines show the expected inferred value in the absence of gene flow for each population (i.e. the mean value for $N_em = 0$). (c) The inferred *difference* in recombination rate between population 1 and population 2 (p_2–p_1) as a function of migration rate. Points and error bars are as in b.

continuous gene flow and secondary contact models. First, in the model of continuous gene flow, when $N_em \geq 1$, we observed a systematic increase (overestimate) in estimated rates of recombination in both populations (Fig. 2, a and b, top row, $N_em = 1$–100). This increase was statistically significant [Type III Wald chi-square = 5090.07, $P < 2.0 \times 10^{-16}$; coefficient for gene flow = 0.63–0.67 (95% CI), $t(19495) = 71.34$, $P < 0.001$]. When the migration rate was moderate to high (N_em 10–100), the recombination rate was overestimated by ∼10–20% (Fig. 2b). This effect is consistent with migration causing the populations to become coupled, behaving as a single population with a larger Ne and thus inflating the population-scaled estimate of recombination rate.

In contrast to the continuous gene flow case, under a model of secondary contact, there was a marked systematic decrease (underestimate) of recombination rates, which also became visible when $N_em \geq 1$ (Fig. 2, a and b, bottom row, $N_em = 1$–100). This decrease was statistically significant [Type III Wald chi-square = 1512, $P < 2.2 \times 10^{-16}$; coefficient for gene flow = −(0.54–0.49) (95% CI), $t(31846) = -38.88$, $P < 0.001$]. The magnitude of this decrease was substantial: on average, populations experiencing $N_em = 1$ had recombination rates about 20% lower than expected, with this increasing to 50% when $N_em = 10$ or higher (Fig. 2, a and b). This decrease was accompanied by a statistically significant increase in the *variance* of recombination rate estimates, especially for $N_em = 1$–10 compared to $N_em < 1$ (Fig. 2a, bottom row; *F*-test for equivalency of variance, $F(10{,}429, 13{,}860) = 0.20863$, $P < 2.2 \times 10^{-16}$). A systematic increase in the mean and variance of LD within populations is consistent with

allele frequency differences between populations manifesting as migration-associated LD, and deflating estimates of recombination rate. When gene flow was very high, there was a visible recovery of estimated recombination rates (Fig. 2a, bottom row, $N_em = 100$), presumably due to migration homogenizing allele frequencies and increased effective population sizes increasing the rate at which recombination breaks down migration-associated LD.

When comparing recombination rates between p_1 and p_2, the "coupling" bias observed in the continuous migration scenario did not appear to systematically affect the *ratio* of recombination rate between the 2 populations (Fig. 2c, Continuous Migration). However, in keeping with the previous result, migration-associated LD in the secondary contact model appeared to greatly increase the between replicate variance in the ratio of recombination rates between populations when $N_em \geq 1$ (Fig. 2c, Secondary Contact).

Inference when recombination rate differs between populations

When recombination rates diverged between populations, we also observed the 2 forms of bias described above (Fig. 3). The estimates from the continuous gene flow scenario exhibited a statistically significant increase [Type III Wald chi-square = 8,936.44, $P < 2.2 \times 10^{-16}$; coefficient for gene flow = 0.65–0.67 (95% CI), $t(19495) = 94.53$, $P < 0.001$] whereas estimates from the secondary contact model exhibited a statistically significant decrease [Type III Wald chi-square = 1,512, $P < 2.0 \times 10^{-16}$; coefficient for gene flow = −(0.27–0.22) (95% CI), $t(34505) = -23.22$, $P < 0.001$]. However, the results differed from simulations with constant recombination rates in a number of important ways. First, there was a clear difference between the continuous migration and secondary contact models in the overall trajectory in the population-specific estimates of recombination rate (Fig. 3a). In the continuous gene flow models, there was an overall positive trend for the estimates of recombination rate in p_2 even in the absence of gene flow (Fig. 3a, continuous migration). This was presumably caused by a lag in the establishment of equilibrium levels of LD within p_2 that reflect the new recombination rate (which spontaneously changed at the time of divergence). This lag resulted in the recombination rate in p_2 being consistently underestimated (because it had not reached its new equilibrium), in addition to the coupling effect observed previously (Fig. 3, b and c, continuous migration). In the case of the secondary contact model, we did not observe the same positive trend for recombination rate estimates in p_2, likely because the isolation period (1.7M generations) was sufficiently long enough for p_2 to establish an equilibrium level of LD prior to secondary contact (Fig. 3a, Secondary Contact).

Additional simulations

To examine the robustness of our results, we explored 2 additional demographic scenarios. First, we repeated our simulations using an effective population size of 1,720 (1/1,000th of the *D. melanogaster* Ne). These simulations produced broadly similar results, with all biases we identified using a larger Ne also appearing in the presence of a smaller Ne (Supplementary Figs. 2 and 3). Furthermore, there was an apparent increase in variance of recombination estimates in many cases, suggesting that the issues we identified may be considerably worse in species with smaller effective population sizes (Supplementary Fig. 2c).

Finally, using these smaller effective population size simulations as a base, we also explored how asymmetry in gene flow may affect the estimation of recombination rate. To do this, we performed simulations in which gene flow is unidirectional from p2 into p1. Asymmetrical gene flow resulted in an increase in both the bias and variance of recombination rates estimates, with the worst effects again manifesting at moderate levels of gene flow (Supplementary Figs. 4 and 5). As such, it appears that our core simulations (large Ne, symmetrical gene flow) may represent a "best case scenario" for mitigating biases, and departures from this balanced scenario appear to only worsen the prospects for accurate population genetic estimation of recombination rates.

In keeping with the scenario with constant recombination rates, starting at $N_em \sim 1$, migration-associated LD resulted in the systematic underestimation and increase in variance for estimated recombination rates within both p_1 and p_2 [Fig. 3b, Secondary Contact; Type III Wald chi-square = 538.97, $P < 2.0 \times 10^{-16}$; coefficient for gene flow = −(0.27–22) (95% CI), $t(34505) = -23.22$, $P < 0.001$; *F*-test for equivalency of variance, $F(7,734, 13,579) = 0.2174$, $P < 2.2 \times 10^{-16}$]. In addition, the observed divergence in recombination rate between p_2 and p_1 (which was always expected to be +2 cM/Mb) decreased with increasing levels of gene flow (Fig. 3, b and c, Secondary Contact). This effect would likely result in an increase in false negatives with increasing gene flow (i.e. finding no difference in recombination rate between populations when there is in fact one). This decrease in the observed divergence between populations is again likely the outcome of the population-specific levels of LD becoming coupled/merged at moderate to high levels of gene flow, resulting in the populations exhibiting LD (and hence recombination rate estimates) intermediate to what would be expected in the absence of gene flow.

Discussion

Accurate estimates of recombination rate are key to understanding the causes and consequences of recombination rate variation in natural populations. With the increasing availability of genome-wide sequencing data, LD-based estimators of recombination rate have become widely used in a large variety of taxa. However, while gene flow is widely known to shape patterns of LD in populations, the effect of gene flow on LD-based estimators of recombination rate remains largely unexplored. Here, we used forward-time simulations to show that (1) moderate to strong gene flow can introduce substantial bias into LD-based estimates of genome-wide recombination rate and (2) the nature of this bias depends on the demographic and evolutionary history of the populations in question.

Our results here are consistent with theoretical predictions that gene flow between populations can affect LD: increasing in the magnitude and variance of LD at low migration rates as well as reducing LD via the "coupling effect" we observed at higher rates of gene flow. Our study shows how these predictions play out with modern methods and genomic data, and also provides a sense of the magnitude of the potential degree of misestimation – in our case, ranging from 20 to 50 percentage points in cases of moderate gene flow. For comparison, a recent study of population-level differences in recombination rate in *Drosophila pseudoobscura* revealed genetically based interpopulation differences on the magnitude of ~10% measured using replicated linkage maps in each population (Samuk *et al.* 2020). Using LD-based estimators, an observed a difference of this magnitude could be spuriously generated by modest levels of gene flow alone, or missed altogether due to coupling at higher levels of gene flow.

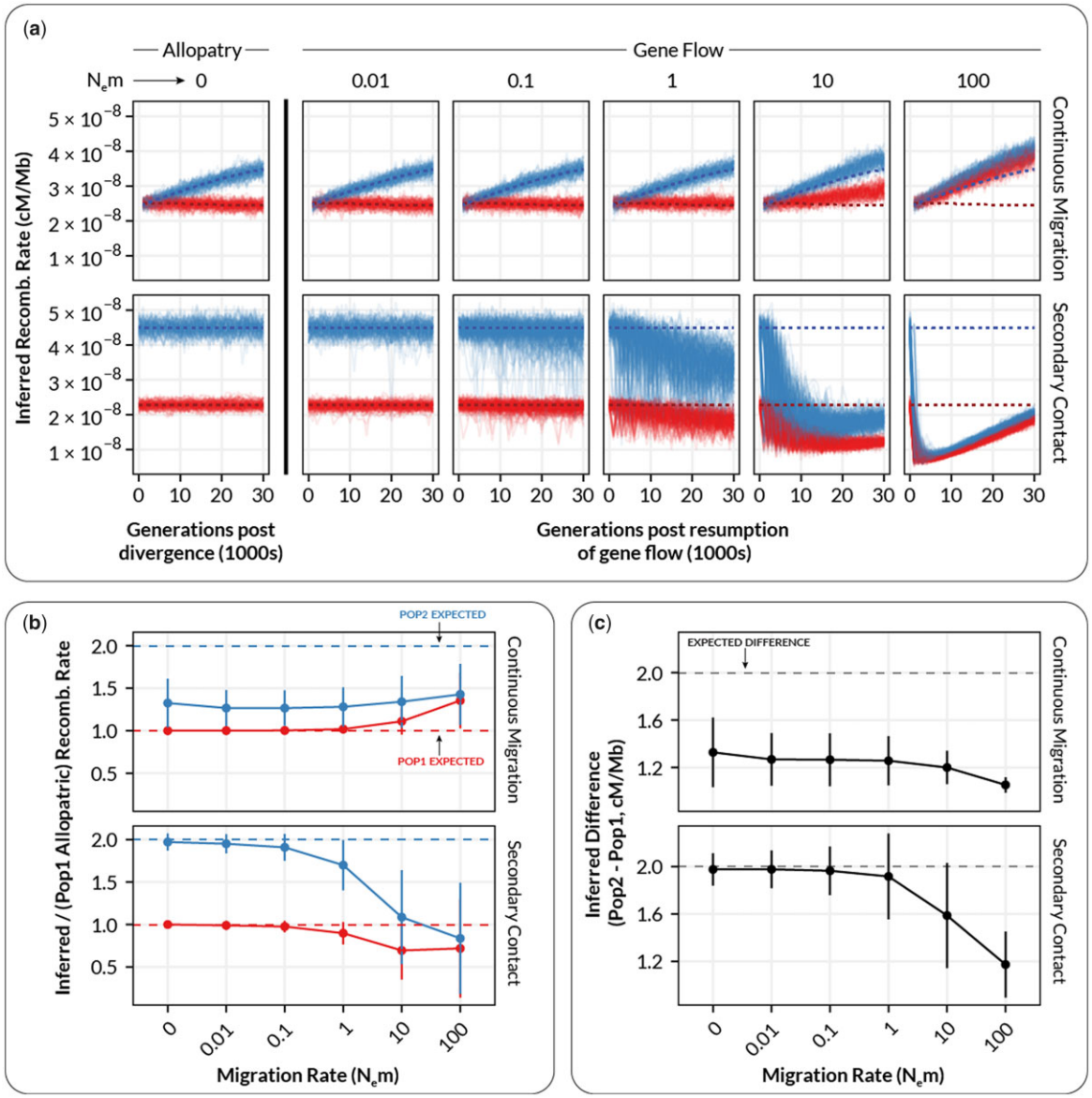

Fig. 3. The relationship between inferred recombination rate and the migration rate in simulated populations where recombination rate increases by a factor of 2 in one subpopulation. (a) Inferred recombination rates for individual simulations at varying levels of migration. Each plot shows inferred rates for simulation replicates (transparent lines) of population 1 (red, unchanged recombination) and population 2 (blue, increased recombination) for a single migration rate. Dashed lines show the expected inferred value in the absence of gene flow (inferred from $N_em = 0$). (b) Summarized inferred recombination rates (y-axis) for each level of migration (x-axis) from the simulations in a. Points are mean values and error bars depict standard deviations (summarized across all generations). Dashed lines show the expected inferred value in the absence of gene flow for each population (i.e. the mean value for $N_em = 0$). (c) The inferred *difference* in recombination rate between population 1 and population 2 ($p2–p_1$) as a function of migration rate. Points and error bars are as in b.

In addition, the specific magnitude and direction of the bias introduced by gene flow are difficult to know without precise knowledge of the population/demographic histories of the populations in question. This should give pause to anyone planning on using LD-based methods to infer recombination rate in nonequilibrium populations.

One key question is whether there are methods to control for or counteract the increased variance and/or biases in the estimation of recombination rate caused by gene flow. One approach could be to identify and remove introgressed haplotypes from datasets prior to inferring recombination rate, thereby removing migration-associated LD. This would require "pure" samples from the source populations, such that the population of origin could be assigned to haplotype blocks (Dias-Alves *et al.* 2018). However, this method would require gene flow to be low enough that coupling (of both LD and allele frequencies) has not occurred. The upward bias and increased variance in recombination rate that occurs as a result of coupling, together with the homogenization of allelic differences between populations at higher levels of gene flow will likely make a "filtering" scheme very difficult (perhaps impossible) to achieve. One other approach may be to attempt to *jointly* estimate a demographic model along with population-specific recombination rates, as has been done with mutation rates (DeWitt *et al.* 2021). However, given the existing complexity and uncertainty in inferring demographic models, we suspect it may be difficult to disentangle the complex interdependencies between gene flow, population size, and estimates of recombination rate.

On a related note, in our simulations we had perfect knowledge of the demographic histories of both populations (ancestral and derived population sizes; divergence time), which was used to parameterize the correction procedure employed by pyrho. In the vast major of empirical cases, demographic history would need to be separately estimated prior to parameterizing pyrho. Such demographic inference is itself error prone and subject to a wide variety of potential biases (Marchi *et al.* 2021) and these errors would propagate into estimates of rho (see Dapper and Payseur, 2018). Furthermore, the interactions between selection, gene flow, and recombination likely further complicate inference of rho. For example, many studies have now shown a negative correlation between recombination rate and introgression mediated by alleles causing reproductive isolation (Aeschbacher *et al.* 2017; Samuk *et al.* 2017; Schumer *et al.* 2018). This suggests that biases in rho estimation introduced by gene flow could themselves vary with genomic context. As such, the biases and error rates identified here represent a "best case" scenario, and would be in *addition* to any errors due to misestimation of the demographic history or the effects of reproductive isolation.

Together with previous work (Dapper and Payseur 2018), our results suggest that LD-based estimates of recombination rate need to be interpreted with great caution when studying nonequilibrium populations. Indeed, these methods are likely only appropriate when populations can be assumed to be evolving in the absence of any gene flow and have reached a reasonable demographic equilibrium. However, it is now widely appreciated that gene flow is ubiquitous in natural populations (Waples and Gaggiotti 2006; Ellstrand and Rieseberg 2016). This may mean that many published LD-based estimates of recombination rate are incorrect. Without empirical maps to compare existing LD-based estimates, it is difficult to say just how incorrect they are. What can be said is that the levels of gene flow required to introduce nontrivial biases into estimates of recombination rate, i.e. N_em ~1–10, are not uncommon in natural populations (Slarkin 1985; Waples and Gaggiotti 2006). Although direct estimates of Nem in wild populations are scarce, under an island model, an Nem of 10 would correspond to an FST of around 0.02. In a review of population comparisons of traits by Leinonen *et al.* (2008), such populations comprise around 20–30% of the cases identified. It is also worth noting that it is not the case that 2 populations being studied have to be exchanging genes themselves (e.g. which would not the case when studying 2 reproductively different species), but just that one or more of the populations are exchanging genes with some *other* population (e.g. an unsampled population of the same species).

If many LD-based estimates are incorrect, why do published LD-based estimates of recombination rate correlate well with direct estimates, e.g. from genetic maps? (McVean and Auton 2007; Chan *et al.* 2012; Smukowski Heil *et al.* 2015). There are several considerations. First, the correlations that have been reported are by no means perfect (e.g. ~Spearman's Rho of 0.6: Smukowski Heil *et al.* 2015; $r^2 = 0.37$–63: (Chan *et al.* 2012) and depend greatly on the genomic scale at which they are measured (Smukowski Heil *et al.* 2015). Second, simple correlations between LD-based and empirical estimates do not speak to genome-wide differences in the estimates of recombination rate, such as those due to the coupling effects we observed. Such effects would be visible as differences in the *intercept* of a linear regression, rather than the R^2, for example. Finally, the species where these correlations have been examined (humans and *D. melanogaster*) may meet the assumptions of demographic equilibrium more readily (Ochoa and Storey 2019; Suvorov *et al.* 2022). While such assumptions may be reasonable for these populations, for which LD-based estimators were originally developed, they are much less likely to hold in many natural populations. Notably, they are likely rarely met in populations that have recently adaptively diverged in the presence of gene flow, which have lately been the subject of increased research interest (Ravinet *et al.* 2017; Linck and Battey 2019). The equilibrium assumption is also likely not valid in populations in which the recombination rate has recently changed (Brandvain and Wright 2016), reducing the utility of these estimates for studying the rapid evolution of recombination rates.

While we only focused on a single implementation of one type of LD-based estimator of recombination (pyrho), it is likely that other population genetic methods will also suffer from the effects we describe here. LD is the "information" used by all estimators, either directly as in methods like LDjump (Hermann *et al.* 2019) or indirectly as in machine learning methods like ReLERNN (Adrion, Galloway *et al.* 2020). That said, in the case of the latter method, it may be possible to overcome some of the issues we have identified if the training datasets were simulated with an accurate demographic model. As such, the distorting effects of gene flow on LD need to be carefully considered when applying any statistical methods for inferring recombination rate approaches. We also stress that our simulations do not suggest that LD-based estimators and their implementations are wrong per se, but rather that the assumptions under which LD-based estimates are biologically accurate are readily violated by levels of gene flow and divergence common seen in natural populations.

Finally, the biases we have identified likely affect the identification of recombination cold/hotspots and the "landscape" of recombination in general. For example, if introgression is itself variable across the genome, this could result in the biases we have identified here (1) covarying with introgression and (2) creating false heterogeneity in recombination estimates. The increase in variance we identified could also result in (apparent) increased heterogeneity in recombination across the genome. In terms of identifying hotspots, the difference in recombination rate between hot and cold spots in most species vastly exceeds the 20–50% differences we described here, and thus the biases we identified may not be an issue for the identification of extreme hotspots per se.

Conclusion

Studying variation in recombination rate is difficult. LD-based methods for inferring recombination rate are attractive in their data requirements but require strong assumptions to be met. As we have shown here, gene flow readily violates these assumptions and introduces biases and decreases in precision, in a variety of ways that are difficult to identify in a given study population. This is problematic because gene flow is extremely common in natural populations. How should we proceed? Rather than attempt to squeeze blood from the proverbial stone, we believe that the most straightforward solution to the problems we outline here is simply to prioritize the use of direct, empirical methods for measuring of recombination rate. This decision is made hopefully simpler with the increased ease and low cost of creating traditional linkage maps and performing gamete sequencing. That said, LD-based approaches remain important tools for hypothesis generation, and when paired with direct estimates of recombination rate can provide a detailed picture of both the past and present landscape of recombination rates in natural populations.

Data availability

All codes used to generate the simulated data used in this study are available as a repository on Github (http://github.com/ksamuk/LD_recomb).

Supplemental material is available at G3 online.

Acknowledgments

The authors thank members of the Noor lab and Dr Katharine Korunes for helpful discussions and for providing comments on an early draft of this article.

Funding

Support this project was provided by the National Science Foundation grants DEB-1545627, 1754022, and 1754439 to MAFN. KS was additionally supported by a Natural Sciences and Engineering Research Council of Canada Postdoctoral Fellowship.

Conflicts of interest

None declared.

Literature cited

Adrion JR, Cole CB, Dukler N, Galloway JG, Gladstein AL, Gower G, Kyriazis CC, Ragsdale AP, Tsambos G, Baumdicker F, *et al.* A community-maintained standard library of population genetic models. Elife. 2020;9:e54967.

Adrion JR, Galloway JG, Kern AD. Predicting the landscape of recombination using deep learning. Mol Biol Evol. 2020;37(6):1790–1808.

Aeschbacher S, Selby JP, Willis JH, Coop G. Population-genomic inference of the strength and timing of selection against gene flow. Proc Natl Acad Sci USA. 2017;114(27):7061–7066.

Auton A, McVean G. Recombination rate estimation in the presence of hotspots. Genome Res. 2007;17(8):1219–1227.

Barton NH. The role of hybridization in evolution. Mol Ecol. 2001; 10(3):551–568.

Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw. 2015;67(1):1–48. doi: 10.18637/jss.v067.i01.

Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature. 1992;356(6369):519–520.

Brandvain Y, Wright SI. The limits of natural selection in a nonequilibrium world. Trends Genet. 2016;32(4):201–210.

Broman KW. Genetic Map Construction with R/Qtl. Technical Report # 214. Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison; 2010.

Burri R. Interpreting differentiation landscapes in the light of long-term linked selection. Evol Lett. 2017;1(3):118–131.

Chan AH, Jenkins PA, Song YS. Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. PLoS Genet. 2012; 8(12):e1003090.

Comeron JM, Ratnappan R, Bailin S. The many landscapes of recombination in *Drosophila melanogaster*. PLoS Genet. 2012;8(10): e1002905.

Cutter AD. Recombination and linkage disequilibrium in evolutionary signatures. A Primer of Molecular Population Genetics. 2019; 113–128.

Dapper AL, Payseur BA. Connecting theory and data to understand recombination rate evolution. Phil Trans R Soc B. 2017;372(1736): 20160469.

Dapper AL, Payseur BA. Effects of demographic history on the detection of recombination hotspots from linkage disequilibrium. Mol Biol Evol. 2018;35(2):335–353.

DeWitt WS, Harris KD, Ragsdale AP, Harris K. Nonparametric coalescent inference of mutation spectrum history and demography. Proc Natl Acad Sci U S A. 2021;118. https://doi.org/10.1073/pnas.2013798118

Dias-Alves T, Mairal J, Blum MGB. Loter: a software package to infer local ancestry for a wide range of species. Mol Biol Evol. 2018; 35(9):2318–2326.

Dréau A, Venu V, Avdievich E, Gaspar L, Jones FC. Genome-wide recombination map construction from single individuals using linked-read sequencing. Nat Commun. 2019;10(1):4309.

Dumont BL, Payseur BA. Evolution of the genomic rate of recombination in mammals. Evolution. 2008;62(2):276–294.

Ellstrand NC, Rieseberg LH. When gene flow really matters: gene flow in applied evolutionary biology. Evol Appl. 2016;9(7): 833–836.

Haddrill PR, Charlesworth B, Halligan DL, Campos JL. The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. Mol Biol Evol. 2014;31(4): 1010–1028.

Haller BC, Messer PW. SLiM 3: forward genetic simulations beyond the Wright-Fisher model. Mol Biol Evol. 2019;36(3):632–637.

Hermann P, Heissl A, Tiemann-Boege I, Futschik A. LDJump: estimating variable recombination rates from population genetic data. Mol Ecol Resour. 2019;19(3):623–638.

Hunter CM, Huang W, Mackay TFC, Singh ND. The genetic architecture of natural variation in recombination rate in *Drosophila melanogaster*. PLoS Genet. 2016;12(4):e1005951.

Johnston SE, Bérénos C, Slate J, Pemberton JM. Conserved genetic architecture underlying individual recombination rate variation in a wild population of soay sheep (*Ovis aries*). Genetics. 2016;203(1):583–598.

Kamm JA, Spence JP, Chan J, Song YS. Two-locus likelihoods under variable population size and fine-scale recombination rate estimation. Genetics. 2016;203(3):1381–1399.

Korunes KL, Samuk K, Noor MAF. Disentangling types of linked selection using patterns of nucleotide variation in *Drosophila pseudoobscura*. In: Population Genomics. Cham: Springer International Publishing; 2021. pp. 1–22.

Leinonen T, O'Hara RB, Cano JM, Merilä J. Comparative studies of quantitative trait and neutral marker divergence: a meta-analysis. J Evol Biol. 2008;21(1):1–17.

Li H, Stephan W. Inferring the demographic history and rate of adaptive substitution in Drosophila. PLoS Genet. 2006;2(10):e166.

Linck E, Battey CJ. On the relative ease of speciation with periodic gene flow. bioRxiv. 2019. 758664. https://doi.org/10.1101/758664.

Lloyd A, Morgan C, H Franklin FC, Bomblies K. Plasticity of meiotic recombination rates in response to temperature in Arabidopsis. Genetics. 2018;208(4):1409–1420.

Mallet J. Hybridization as an invasion of the genome. Trends Ecol Evol. 2005;20(5):229–237.

Marchi N, Schlichta F, Excoffier L. Demographic inference. Curr Biol. 2021;31(6):R276–R279.

McVean G. Linkage disequilibrium, recombination and selection. In: DJ Balding, M Bishop, C Cannings, editors. Handbook of Statistical Genetics. Chichester (UK): John Wiley & Sons, Ltd; 2007. pp. 909–944.

McVean G, Auton A. LDhat 2.1: A Package for the Population Genetic Analysis of Recombination. Oxford (UK): Department of Statistics; 2007.

Morjan CL, Rieseberg LH. How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. Mol Ecol. 2004;13(6):1341–1356.

Nei M, Li W-H. Linkage disequilibrium in subdivided populations. Genetics. 1973;75(1):213–219.

Ochoa A, Storey JD. New kinship and FST estimates reveal higher levels of differentiation in the global human population. BioRxiv. 2019. https://doi.org/10.1101/653279.

Ohta T. Linkage disequilibrium with the island model. Genetics. 1982;101(1):139–155.

Peñalba JV, Wolf JBW. From molecules to populations: appreciating and estimating recombination rate variation. Nat Rev Genet. 2020;21(8):476–492.

Peterson AL, Miller ND, Payseur BA. Conservation of the genome-wide recombination rate in white-footed mice. Heredity. 2019; 123(4):442–457.

Peterson AL, Payseur BA. Sex-specific variation in the genome-wide recombination rate. Genetics. 2021;217(1):1–11.

R Core Team. R: A Language and Environment for Statistical Computing. Vienna (Austria): R Foundation for Statistical Computing; 2018.

Rastas P. Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. Bioinformatics. 2017;33(23): 3726–3732.

Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlović M, Noor MAF, Mehlig B, Westram AM. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. J Evol Biol. 2017;30(8):1450–1477.

Ritz KR, Noor MAF, Singh ND. Variation in recombination rate: adaptive or not? Trends Genet. 2017;33(5):364–374.

Rommel Fuentes R, Hesselink T, Nieuwenhuis R, Bakker L, Schijlen E, Dooijeweert W, Diaz Trivino S, Haan JR, Sanchez Perez G, Zhang X, *et al.* Meiotic recombination profiling of interspecific hybrid F1 tomato pollen by linked read sequencing. Plant J. 2020; 102(3):480–492.

Roux C, Fraïsse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. Shedding light on the grey zone of speciation along a continuum of genomic divergence. PLoS Biol. 2016;14(12):e2000234.

Samuk K, Manzano-Winkler B, Ritz KR, Noor MAF. Natural selection shapes variation in genome-wide recombination rate in *Drosophila pseudoobscura*. Curr Biol. 2020;30(8):1517–1528.e6.

Samuk K, Owens GL, Delmore KE, Miller SE, Rennison DJ, Schluter D. Gene flow and selection interact to promote adaptive divergence in regions of low recombination. Mol Ecol. 2017;26(17):4378–4390.

Schumer M, Xu C, Powell DL, Durvasula A, Skov L, Holland C, Blazier JC, Sankararaman S, Andolfatto P, Rosenthal GG, *et al.* Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science. 2018;360(6389):656–660.

Slarkin M. Gene flow in natural populations. Annu Rev Ecol Syst. 1985;16(1):393–430.

Smukowski Heil CS, Ellison C, Dubin M, Noor MAF. Recombining without hotspots: a comprehensive evolutionary portrait of recombination in two closely related species of Drosophila. Genome Biol Evol. 2015;7(10):2829–2842.

Spence JP, Song YS. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. Sci Adv. 2019;5(10):eaaw9206.

Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Philos Trans R Soc Lond B Biol Sci. 2017;372:20160455.

Stumpf MPH, McVean GAT. Estimating recombination rates from population-genetic data. Nat Rev Genet. 2003;4(12):959–968.

Suvorov A, Kim BY, Wang J, Armstrong EE, Peede D, D'Agostino ERR, Price DK, Waddell P, Lang M, Courtier-Orgogozo V, *et al.* Widespread introgression across a phylogeny of 155 Drosophila genomes. Curr Biol. 2022;32(1):111–123.e5.

Tiley GP, Burleigh JG. The relationship of recombination rate, genome structure, and patterns of molecular evolution across angiosperms. BMC Evol Biol. 2015;15(1):194.

Waples RS, Gaggiotti O. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol Ecol. 2006;15(6): 1419–1439.

Wickham H. The tidyverse. R package ver. 1.1. 2017.

Xu P, Kennell T, Gao M, Kimberly RP, Chong Z; Human Genome Structural Variation Consortium. MRLR: unraveling high-resolution meiotic recombination by linked reads. Bioinformatics. 2020;36(1):10–16.

Communicating editor: J. Ross-Ibarra