Gene flow biases population genetic inference of recombination rate

K. Samuk[1,2]* & M.A.F. Noor[1]

[1] Department of Biology, Duke University.

[2] Department of Evolution, Ecology, and Organismal Biology, University of California, Riverside

*Corresponding author, 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 recombination rate itself. Model-based population genetic methods that infer recombination rates from patterns of linkage disequilibrium (LD) in the genome have become a popular method to estimate rates of recombination. However, these LD-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 LD-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 positive 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.

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 & Aquadro,

1992; Burri, 2017; Cutter, 2019; Haddrill et al., 2014; Korunes et al., 2021). Genome-wide rates of recombination also modulate diverse processes such as adaptation, speciation, and introgression (Dapper & Payseur, 2017; Samuk et al., 2017; Schumer et al., 2018; Stapley et al., 2017). There is also a growing appreciation that recombination rate is itself a trait that varies and evolves (Dumont & Payseur, 2008; Hunter et al., 2016; Johnston et al., 2016; Ritz et al., 2017; Samuk et al., 2020; Stapley et al., 2017). Accordingly, there has been great interest in efficient and accurate methods for estimating recombination rates.

Current methods for estimating recombination rates fall into two broad classes of methods: direct and indirect (Peñalba & Wolf, 2020). Of the direct measures, the three 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 carry out in the laboratory or unavailable in some populations. Further, 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., 2019; 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 & Payseur, 2021). However, 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 & McVean, 2007; Chan, Song, et al., 2012; Kamm et al., 2016; Spence & Song, 2019). These methods attempt to estimate recombination rates by statistically fitting recombination rates (derived from population genetic models/simulations) to observed 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 & Song, 2019) and (3) are informative of time-averaged population historical recombination rates (Gil McVean & Auton, 2007). Accordingly, LD-based estimates of recombination rates have become extremely popular, and now vastly outnumber direct measures in the literature (Peñalba & Wolf, 2020; Stapley et al., 2017). These methods have also begun to be used to perform interpopulation comparisons of recombination rates (Peñalba & Wolf, 2020; Stapley et al., 2017).

3

68   Like all models, LD-based estimators of recombination rate make a variety of

69   simplifying assumptions about the populations of interest. For one, they generally

70   assume that the populations/loci of interest are evolving largely neutrally and have

71   reached population genetic equilibrium in a number of ways (Stumpf & McVean, 2003).

72   In particular, most methods assume that the populations being studied have reached an

73   equilibrium between recombination and population scaled mutation, such that LD

74   accurately reflects patterns of recombination rate (G. McVean, 2007). Further, it is

75   generally assumed that any form of selection that might distort patterns of LD (e.g.

76   sweeps) has not recently occurred (Chan, Song, et al., 2012). Finally, these methods make

77   the general assumption that demographic processes that distort genome-wide patterns

78   of LD, such as population size changes, have not occurred (recall that ρ is directly

79   dependent on N_e, Auton & McVean, 2007).

80   While some of these assumptions may be robust to violation, work has shown that some

81   violations can result in biased estimates. For example, (Dapper & Payseur, 2018) showed

82   that recombination estimates from LDhat (Gil McVean & Auton, 2007) are highly

83   sensitive to changes in population size. This can be ameliorated in some cases by

84   incorporating known changes in population size into the estimation procedure, such as

85   implemented in the software pyrho (Spence & Song, 2019).

86   Along with changes in population size and selection, another process that can greatly

87   alter patterns of LD is gene flow. Gene flow and subsequent admixture between

88   diverged populations can have complex effects on patterns of LD within each

89   population (Nei & Li, 1973; Ohta, 1982). These effects range from large and genomically

90   variable increases in LD due to segregation of divergent haplotypes, to genome-wide

91   decreases in LD as populations become coupled and increase local N_e (Nei & Li, 1973;

92   Ohta, 1982). While it is now widely accepted that gene flow is commonplace in natural

93   populations (Barton, 2008; Mallet, 2005; Suvorov et al., 2021; Waples & Gaggiotti, 2006),

94   and there has not been a systematic study of the effects of gene flow on LD-based

95   measures of recombination. Further, it remains unclear how gene flow (or any other

96   violation of assumptions) impacts our ability to detect differences in recombination rate

97   *between* (as opposed to within) populations using LD-based methods.

98   Here, we address these issues using forward-time population genetic simulations. We

99   attempt to answer two specific questions. First, how does gene flow between

100   populations affect the precision and accuracy of LD-based estimates of recombination

101   rate within populations? Secondly, how does gene flow affect our ability to detect

102   evolved differences in recombination rate between populations? Our primary goal is to

103   answer these questions in the context of a core set of realistic demographic scenarios,

104   and not perform an exhaustive exploration of parameter space. Overall we hope to help

105   investigators understand key sources of bias in LD-based estimates of recombination

106   rate in natural populations and highlight areas of future development.

Methods

107

Code availability

108

109   All scripts used in the analyses described below are available as a repository on Github

110   (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 & Messer, 2019). The basic form of all the simulations was an isolation-with-migration scenario: a single ancestral population diverges into two subpopulations with a static amount of bidirectional gene flow (Figure 1). Populations were composed of diploid individuals with 100kb 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)*: Mutation rate = 5.49×10^{-9} (Li & Stephan, 2006); Recombination rate = 2.23×10^{-8}, (average of chromosome 2R, Comeron et al., 2012); N_e = 1.72M (Li & Stephan, 2006). Recombination and mutation rates were conservatively modeled as uniform across the 100kb genome. Following standard practice for forward-time simulations, all simulations were run with an *in silico* population size of N=1000, and simulated mutation and recombination rates scaled by a factor of N_e/N as per the SLiM manual (Haller & 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 ≈ 1751 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 five orders of magnitude: 0, 0.01, 0.1, 1, 10, 100, in standard units of $N_e m$ (the product of the effective population size and the migration rate). These

6

134   values were chosen to encompass total isolation ($N_em = 0$), limited gene flow ($N_em = $

135   0.01-0.1), moderate gene flow in interconnected metapopulations ($N_em = 1\text{-}10$, Morjan &

136   Rieseberg, 2004; Waples & Gaggiotti, 2006), and a scenario of a nascent hybrid swarm

137   ($N_em = 100$). We also varied the timing of the onset of gene flow, with gene flow

138   beginning either immediately after divergence or after a period of isolation. We

139   performed preliminary simulations to determine a period of isolation (~1.7M

140   generations in our case) that produced levels of genomic divergence (Figure S1) similar

141   to those observed in natural population pairs that exhibit genome-wide genetic

142   divergence but still actively exchange genes ($F_{ST} \sim 0.4$, Morjan & Rieseberg, 2004; Roux

143   et al., 2016). Finally, to explore how gene flow impacts the detection of population

144   *differences* in recombination rate, we modeled scenarios where recombination rate either

145   remains constant in both subpopulations or instantaneously increases by a factor of 2 at

146   the time of divergence in one of the two subpopulations (always subpopulation two).

147   This magnitude of this difference is well within the range of variation in recombination

148   rate reported for a wide variety of species (Stapley et al., 2017). In biological terms, an

149   instantaneous increase in population recombination rate could be readily mediated by

150   an environmental change (e.g. temperature, Lloyd et al., 2018) or via a change in mating

151   system (Brandvain & Wright, 2016). We note that this instantaneous change is a "best

152   case" scenario for detecting interpopulation differences in recombination rate, and thus

153   any loss of power to detect differences in recombination that occurs due to gene flow

154   will be conservative.
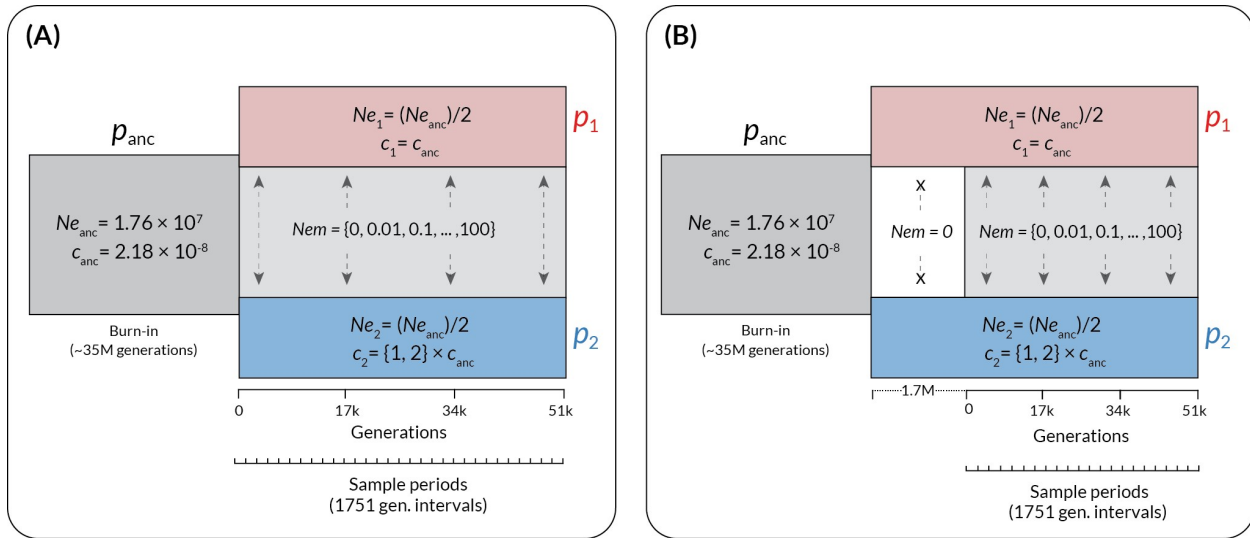
Figure 1 | The structure of the forward-time simulations performed in SLiM. Time in back-transformed generations is shown along the x-axis, and the populations in existence at a given time are shown as rectangles. p_{anc} = the ancestral population, p_1 = the subpopulation with unchanged recombination 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 post-divergence is indicated below the plots, with the pre-contact isolation period in (B) shown as a dotted line preceding the main axis. Sample periods indicate intervals at which genotypes were output for analysis.

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 & Messer, 2019). This initial period was followed by divergence into two 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

8

171   occurred at the time of divergence and instantaneously applied to all individuals in

172   subpopulation two only.

173   Starting at the time of divergence and thereafter in intervals of 1751 generations, we

174   collected a random sample of 25 individuals (a total of 50 haploid genomes) from each

175   population and saved their complete genotypic at all sites in VCF format. We stopped

176   the simulations after 51 000 generations. Each parameter combination was replicated

177   100 times, for a total number of ~n=48 000 population samples.

178   Estimation of recombination rate using pyrho

179   While there are a variety of LD-based estimators of recombination rate, we elected to

180   use pyrho (Spence & Song, 2019) for estimation in this study. It shares its statistical

181   foundation with the most widely used LD-based estimators (LD-hat & LD-helmet; Chan,

182   Jenkins, et al., 2012; Gil McVean & Auton, 2007) while also having the ability to account

183   for changes in effective population size such as we are modeling here (Spence & Song,

184   2019). As such, any estimation biases caused by gene flow will likely affect those

185   approaches as least as much they affect pyrho. Direct comparisons with other methods

186   are complicated by the fact that pyrho is the only model-based method that adequately

187   accounts for changes in effective population (ReLERNN being a possible exception

188   Adrion, Galloway, et al., 2020)

189   We followed the recommended practices for inferring recombination rate using pyrho

190   (https://github.com/popgenmethods/pyrho). We parameterized the initial lookup tables

191   using the effective population size and mutation rates used in the simulations (*unscaled*

192   in this case). To account for changes in effective population size, we created lookup

193   tables that accounted for a change of $N_e/2$ (1.72M to 8.6M) in time steps of 1751

194   generations in the past. This allowed us to have an appropriately timed lookup table for

195   each step of the simulation. We used the built-in methods to infer the hyperparameters

196   of window size (best fit 100) and block penalty (best fit 1000). Using this baseline, we

197   inferred recombination rates using the genotype data (VCF format) from both

198   subpopulations at each time point, for a total of ~96 000 pyrho fits. All computation was

199   performed using the Duke University Computing Cluster, running CentOS Version 8.

200   Statistical analyses

201   We performed all data processing and visualization using the tools of the tidyverse

202   package in R 4.0.3 (R Core Team, 2018; Wickham, 2017). To examine how gene flow

203   between populations affects the accuracy of LD-based estimates of recombination, and

204   the context of the various factors explored in our simulations, we performed an analysis

205   of variance using a linear mixed model with Gaussian errors fitted via the lmer()

206   function from the lme4 package (Bates et al., 2007). This model had the following form:

207   Recombination rate = (1|simulation replicate) + (1|simulation generation) + gene flow

208   magnitude + recombination rate change, where (1|[factor]) denotes a random intercept

209   and ":" denotes an interaction. All variables were standardized (mean-centered and

210   scaled by standard deviation) prior to analysis. To simplify interpretation, we fitted

211   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 two types of systematic biases in estimates of recombination rate within populations (Figure 2A). These effects began when $N_em >= 1$ in both the continuous gene flow and secondary contact models. First, in the model of continuous gene flow, when $N_em >= 1$, we observed a systematic increase (overestimate) in estimated rates of recombination in both populations (Figure 2A, 2B, top row, $N_em = 1-100$). This increase was statistically significant (Type III Wald chi-square = 5090.07, p < 2.0×10^{-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 approximately 10-20% (Figure 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 $N_em >= 1$ (Figure 2A, 2B, bottom row, $N_em = 1-100$). This decrease was statistically significant (Type III Wald chi-square = 1512, p < 2.2×10^{-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

11

234    $N_e m$ = 10 or higher (Figure 2A, Figure 2B). This decrease was accompanied by a

235   statistically significant increase in the *variance* of recombination rate estimates,

236   especially for $N_e m$ = 1-10 compared to $N_e m$ < 1 (Fig 2A, bottom row; F-test for

237   equivalency of variance, $F(10429,13860) = 0.20863$, $p < 2.2 \times 10^{-16}$). A systematic increase in

238   the mean and variance of LD within populations is consistent with allele frequency

239   differences between populations manifesting as migration-associated LD, and deflating

240   estimates of recombination rate. When gene flow was very high, there was a visible

241   recovery of estimated recombination rates (Figure 2A, bottom row, $N_e m$ = 100),

242   presumably due to migration homogenizing allele frequencies and increased effective

243   population sizes increasing the rate at which recombination breaks down migration-

244   associated LD.

245   When comparing recombination rates between p_1 and p_2, the "coupling" bias observed

246   in the continuous migration scenario did not appear to systematically affect the *ratio* of

247   recombination rate between the two populations (Figure 2C, Continuous Migration).

248   However, in keeping with the previous result, migration-associated LD in the secondary

249   contact model appeared to greatly increase the variance in the ratio of recombination

250   rates between populations when $N_e m >= 1$ (Figure 2C, Secondary Contact).

251   Inference when recombination rate differs between populations

252   When recombination rates diverged between populations, we also observed the two

253   forms of bias described above (Figure 3). The estimates from the continuous gene flow

254   scenario exhibited a statistically significant increase (Type III Wald chi-square =

255   8936.44, $p < 2.2 \times 10^{-16}$; coefficient for gene flow =0.65-0.67 (95% CI), $t(19495) = 94.53$, p <

256 0.001) whereas estimates from the secondary contact model exhibited a statistically

257 significant decrease (Type III Wald chi-square = 1512, p < 2.0×10^{-16}; coefficient for gene

258 flow = -(0.27-0.22) (95% CI), t(34505) = -23.22, p < 0.001). However, the results differed

259 from simulations with constant recombination rates in a number of important ways.

260 First, there was a clear difference between the continuous migration and secondary

261 contact models in the overall trajectory in the population-specific estimates of

262 recombination rate (Figure 3A). In the continuous gene flow models, there was an

263 overall positive trend for the estimates of recombination rate in p_2 even in the absence

264 of gene flow (Figure 3A, continuous migration). This was presumably caused by a lag in

265 the establishment of equilibrium levels of LD within p_2 that reflect the new

266 recombination rate (which spontaneously changed at the time of divergence). This lag

267 resulted in the recombination rate in p_2 being consistently underestimated (because it

268 had not reached its new equilibrium), in addition to the coupling effect observed

269 previously (Figure 3B and 3C, continuous migration).In the case of the secondary

270 contact model, we did not observe the same positive trend for recombination rate

271 estimates in p_2, likely because the isolation period (1.7M generations) was sufficiently

272 long enough for p_2 to establish an equilibrium level of LD prior to secondary contact

273 (Figure 3A, Secondary Contact).
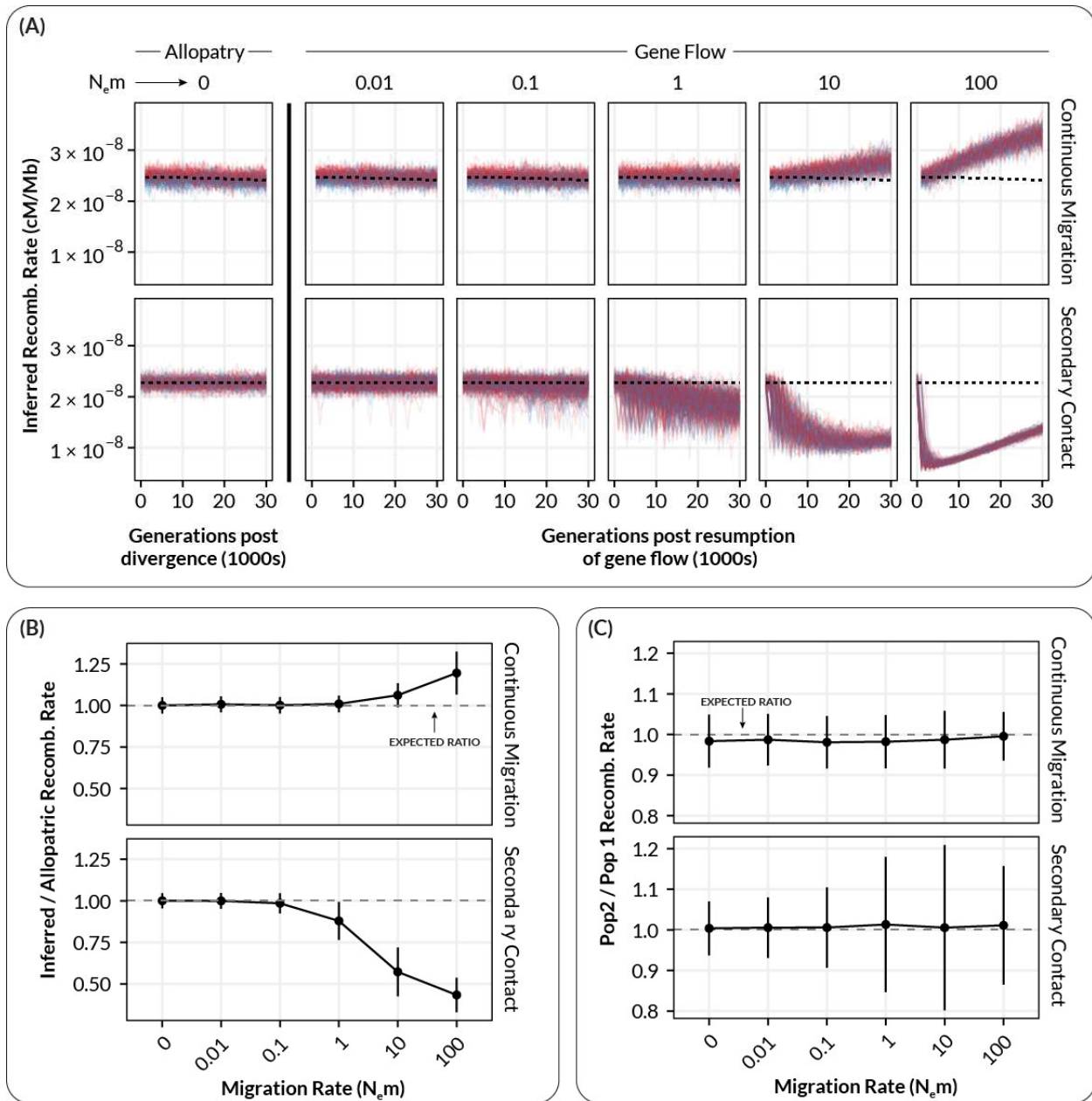
Figure 2 | The relationship between inferred recombination rate and the migration rate in simulated populations where recombination rate remains constant in both subpopulations. (A) Inferred recombination rates for individual simulations at varying levels of migration. Each plot shows inferred rates 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

14

281   error bars depict standard deviations (summarized across all generations). Dashed lines show the

282   expected inferred value in the absence of gene flow for each population (i.e. the mean value for N_em = 0).

283   (C) The inferred *difference* in recombination rate between population 1 and population 2 (p_2 - p_1) as a

284   function of migration rate. Points and errors bars are as in B.

285   In keeping with the scenario with constant recombination rates, starting at N_em ~ 1,

286   migration-associated LD resulted in the systematic underestimation and increase in

287   variance for estimated recombination rates within both p_1 and p_2 (Figure 3B, Secondary

288   Contact; Type III Wald chi-square = 538.97, p < 2.0×10^{-16}; coefficient for gene flow = -

289   (0.27-22) (95% CI), t(34505) = -23.22, p < 0.001; F-test for equivalency of variance,

290   F(7734,13579) = 0.2174, p< 2.2×10^{-16}). In addition, the observed divergence in

291   recombination rate between p_2 and p_1 (which was always expected to be +2 cM/Mb)

292   decreased with increasing levels of gene flow (Figures 3B and 3C, Secondary Contact).

293   This effect would likely result in an increase in false negatives with increasing gene

294   flow (i.e. finding no difference in recombination rate between populations when there is

295   in fact one). This decrease in the observed divergence between populations is again

296   likely the outcome of the population-specific levels of LD becoming coupled/merged at

297   moderate to high levels of gene flow, resulting in the populations exhibiting LD (and

298   hence recombination rate estimates) intermediate to what would be expected in the

299   absence of gene flow.

15
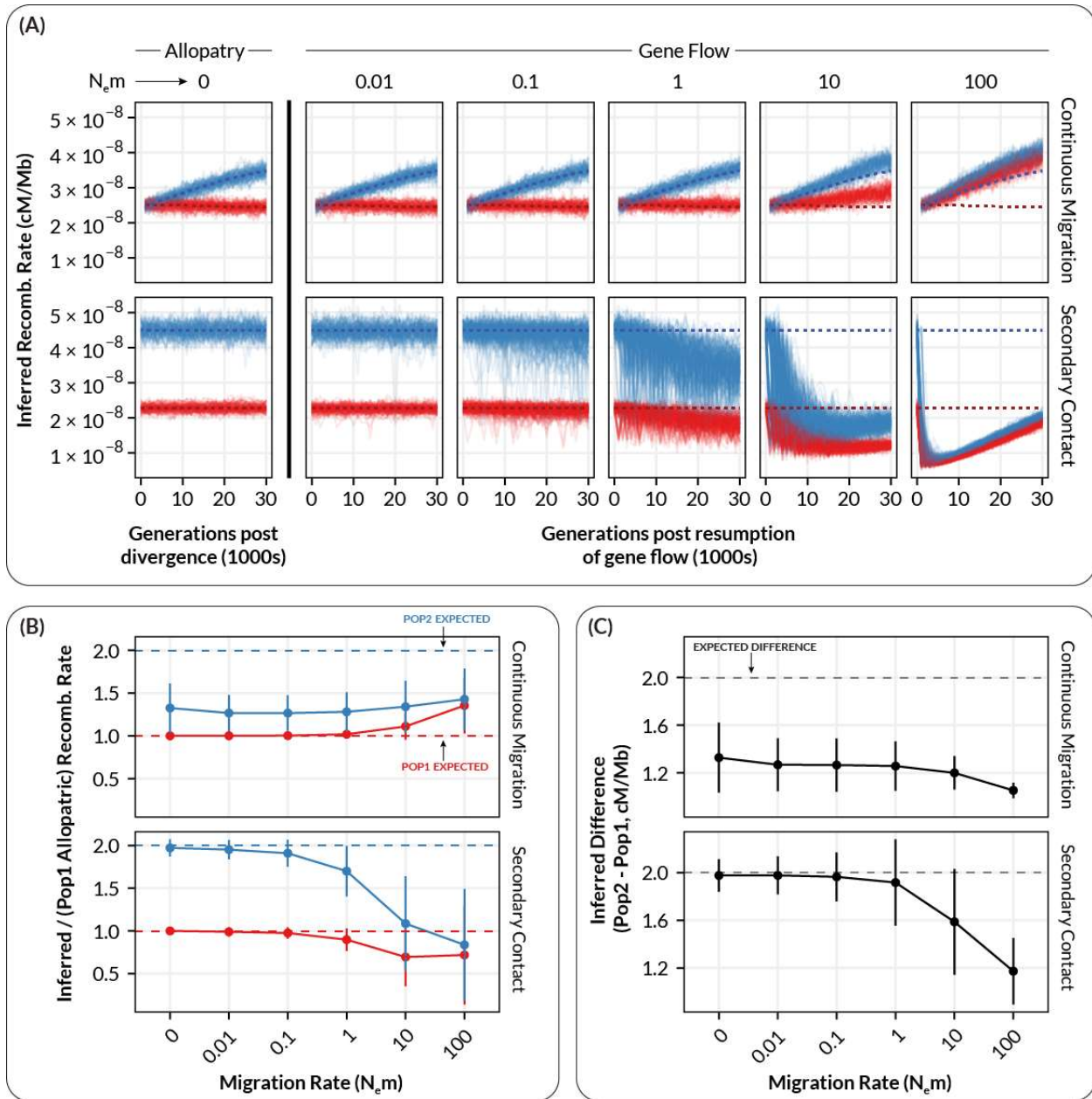
Figure 3 | The relationship between inferred recombination rate and the migration rate in simulated populations where recombination rate increases by a factor of two in one subpopulation. (A) Inferred recombination rates for individual simulations at varying levels of migration. Each plot shows inferred 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

307 error bars depict standard deviations (summarized across all generations). Dashed lines show the

308 expected inferred value in the absence of gene flow for each population (i.e. the mean value for N_em = 0).

309 (C) The inferred *difference* in recombination rate between population 1 and population 2 (*p2* - *p1*) as a

310 function of migration rate. Points and errors bars are as in B.

Discussion

312 Accurate estimates of recombination rate are key to understanding the causes and

313 consequences of recombination rate variation in natural populations. With the

314 increasing availability of genome-wide sequencing data, LD-based estimators of

315 recombination rate have become widely used in a large variety of taxa. However, while

316 gene flow is widely known to shape patterns of LD in populations, the effect of gene

317 flow on LD-based estimators of recombination rate remains largely unexplored. Here,

318 use forward-time simulations to show that (1) gene flow can introduce substantial bias

319 into LD-based estimates of genome-wide recombination rate and (2) the nature of this

320 bias depends on the demographic and evolutionary history of the populations in

321 question.

322 Our results here are consistent with theoretical predictions that gene flow between

323 populations can affect LD: increasing in the magnitude and variance of LD at low

324 migration rates as well as reducing LD via the "coupling effect" we observed at higher

325 rates of gene flow. Our study shows how these predictions play out with modern

326 methods and genomic data, and also provides a sense of the magnitude of the potential

327 degree of misestimation - in our case, ranging from 20-50 percentage points in cases of

328 moderate gene flow. For comparison, a recent study of population-level differences in

329   recombination rate in *Drosophila pseudoobscura* revealed genetically based

330   interpopulation differences on the magnitude of approximately 10% measured using

331   replicated linkage maps in each population (Samuk et al., 2020). Using LD-based

332   estimators, an observed difference of this magnitude could be spuriously generated by

333   modest levels of gene flow alone, or missed altogether due to coupling at higher level

334   gene flow is high. In addition, the specific magnitude and direction of the bias

335   introduced by gene flow is difficult to know without precise knowledge of the

336   population/demographic histories of the populations in question. This should give

337   pause to anyone planning on using LD-based methods to infer recombination rate in

338   non-equilibrium populations.

339   One key question is whether there are methods to control for or counteract the

340   increased variance and/or biases in the estimation of recombination rate caused by gene

341   flow. One approach could be to identify and remove introgressed haplotypes from

342   datasets prior to inferring recombination rate, thereby removing migration-associated

343   LD. This would require "pure" samples from the source populations, such that the

344   population of origin could be assigned to haplotype blocks (Dias-Alves et al., 2018).

345   However, this method would only work if gene flow is infrequent enough that coupling

346   (of both LD and allele frequencies) has not occurred. The upward bias and increased

347   variance in recombination rate that occurs as a result of coupling, together with the

348   homogenization of allelic differences between populations at higher levels of gene flow

349   will likely make a "filtering" scheme very difficult (perhaps impossible) to achieve. One

350   approach may be to attempt to jointly estimate a demographic model along with

351   population-specific recombination rates, as has been done with mutation rates (DeWitt

18

352 et al., 2021). However, given the existing complexity and uncertainty in inferring

353 demographic models, we suspect it may be difficult to disentangle the complex

354 interdependencies between gene flow, population size, and estimates of recombination

355 rate.

356 Together with previous work (Dapper & Payseur, 2018), our results suggest that LD-

357 based estimates of recombination rate need to be interpreted with great caution when

358 studying non-equilibrium populations. Indeed, these methods are likely only

359 appropriate when populations can be assumed to be evolving in the absence of any gene

360 flow, and have reached a reasonable demographic equilibrium. However, it is now

361 widely appreciated that gene flow is ubiquitous in natural populations (Ellstrand &

362 Rieseberg, 2016; Waples & Gaggiotti, 2006). This may mean that many published LD-

363 based estimates of recombination rate are incorrect. Without empirical maps to

364 compare existing LD-based estimates, it is difficult to say just how incorrect. What can

365 be said is that the levels of gene flow required to introduce non-trivial biases into

366 estimates of recombination rate, i.e. $N_e m$ ~1-10, are not uncommon in natural

367 populations (Slarkin, 1985; Waples & Gaggiotti, 2006). It is also worth noting that it is

368 not the case that two populations being studied have to be exchanging genes themselves

369 (e.g. which would not the case when studying two reproductively different species), but

370 just that one or more of the populations are exchanging genes with some *other*

371 population (e.g. an unsampled population of the same species).

372 If many LD-based estimates are incorrect, why do published LD-based estimates of

373 recombination rate correlate well with direct estimates, e.g. from genetic maps? (Chan,

374   Jenkins, et al., 2012; Gil McVean & Auton, 2007; Smukowski Heil et al., 2015). There are

375   several considerations. First, the correlations that have been reported are by no means

376   perfect (e.g. ~Spearman's Rho of 0.6: Smukowski-Heil et al. 2015; r^2 = 0.37-63: Chan,

377   Song, et al., 2012) and depend greatly on the genomic scale at which they are measured

378   (Smukowski-Heil et al. 2015). Second, simple correlations between LD-based and

379   empirical estimates cannot detect genome-wide differences in the estimates of

380   recombination rate, such as those due to the coupling effects we observed. Such effects

381   would be visible as differences in the *intercept* of a linear regression, rather than the R^2,

382   for example. Finally, the species where these correlations have been examined (humans

383   and *Drosophila melanogaster*) may meet the assumptions of demographic equilibrium

384   more readily (Ochoa & Storey, 2019; Suvorov et al., 2021). While such assumptions may

385   be reasonable for these populations, for which LD-based estimators were originally

386   developed, they are much less likely to hold in many natural populations. Notably, they

387   are likely rarely met in populations that have recently adaptively diverged in the

388   presence of gene flow, which have lately been the subject of increased research interest

389   (Linck & Battey, 2019; Ravinet et al., 2017). The equilibrium assumption is likely not

390   valid in populations in which the recombination rate has recently changed (Brandvain &

391   Wright, 2016), reducing the utility of these estimates for studying the rapid evolution of

392   recombination rates.

393   While we only focused on a single implementation of one type of LD-based estimator of

394   recombination (pyrho), it is likely that other population genetic methods will also suffer

395   from the effects we describe here. LD is the "information" used by all estimators, either

396   directly as in methods like LDjump (Hermann et al., 2019) or indirectly as in machine

397   learning methods like ReLERNN (Adrion, Galloway, et al., 2020). That said, in the case

398   of the latter method, it may be possible to overcome some of the issues we've identified

399   if the training datasets were simulated with an accurate demographic model. As such,

400   the distorting effects of gene flow on LD need to be carefully considered when applying

401   any statistical methods for inferring recombination rate approaches. We also stress that

402   our simulations do not suggest that LD-based estimators and their implementations are

403   wrong per se, but rather that the assumptions under which LD-based estimates are

404   biologically accurate are readily violated by levels of gene flow and divergence common

405   seen in natural populations.

406   Conclusion

407   Studying variation in recombination rate is difficult. LD-based methods for inferring

408   recombination rate are attractive in their data requirements, but require strong

409   assumptions to be met. As we have shown here, gene flow readily violates these

410   assumptions and introduces biases and decreases in precision, in a variety of ways that

411   are difficult to identify in a given study population. This is problematic because gene

412   flow is extremely common in natural populations. How should we proceed? Rather than

413   attempt to squeeze blood from the proverbial stone, we believe that the most

414   straightforward solution to the problems we outline here is simply to prioritize the use

415   of direct, empirical methods for measuring of recombination rate. This decision is made

416   hopefully simpler with the increased ease and low cost of creating traditional linkage

417   maps and performing gamete sequencing. That said, LD-based approaches remain

418   important tools for hypothesis generation, and when paired with direct estimates of

419 recombination rate can provide a detailed picture of both the past and present

420 landscape of recombination rates in natural populations.

Acknowledgements

References

428 Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G.,

429 Kyriazis, C. C., Ragsdale, A. P., Tsambos, G., Baumdicker, F., Carlson, J.,

430 Cartwright, R. A., Durvasula, A., Gronau, I., Kim, B. Y., McKenzie, P., Messer, P.

431 W., Noskova, E., Ortega-Del Vecchyo, D., … Kern, A. D. (2020). A community-

432 maintained standard library of population genetic models. *eLife*, *9*.

433 https://doi.org/10.7554/eLife.54967

434 Adrion, J. R., Galloway, J. G., & Kern, A. D. (2020). Predicting the Landscape of

435 Recombination Using Deep Learning. *Molecular Biology and Evolution*, *37*(6), 1790–

436 1808.

437 Auton, A., & McVean, G. (2007). Recombination rate estimation in the presence of

438 hotspots. *Genome Research*, *17*(8), 1219–1227.

439 Barton, N. H. (2008). The role of hybridization in evolution. *Molecular Ecology*, *10*(3), 551–

440 568.

441   Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). The lme4 package. *R Package*

442   *Version*, *2*(1), 74.

443   Begun, D. J., & Aquadro, C. F. (1992). Levels of naturally occurring DNA polymorphism

444   correlate with recombination rates in D. melanogaster. *Nature*, *356*(6369), 519–520.

445   Brandvain, Y., & Wright, S. I. (2016). The Limits of Natural Selection in a

446   Nonequilibrium World. *Trends in Genetics: TIG*, *32*(4), 201–210.

447   Broman, K. W. (2010). Genetic map construction with R/qtl. *University of Wisconsin-*

448   *Madison, Department of Biostatistics & Medical*.

449   https://biostat.wisc.edu/~kbroman/publications/tr_214.pdf

450   Burri, R. (2017). Interpreting differentiation landscapes in the light of long-term linked

451   selection. *Evolution Letters*, *1*(3), 118–131.

452   Chan, A. H., Jenkins, P. A., & Song, Y. S. (2012). Genome-wide fine-scale recombination

453   rate variation in Drosophila melanogaster. *PLoS Genetics*, *8*(12), e1003090.

454   Chan, A. H., Song, Y. S., & Jenkins, P. A. (2012). Genome-Wide Fine-Scale

455   Recombination Rate Variation in Drosophila melanogaster. *PLoS Genetics*, *8*(12),

456   e1003090.

457   Comeron, J. M., Ratnappan, R., & Bailin, S. (2012). The many landscapes of

458   recombination in Drosophila melanogaster. *PLoS Genetics*, *8*(10), e1002905.

459   Cutter, A. D. (2019). Recombination and linkage disequilibrium in evolutionary

460   signatures. In *A Primer of Molecular Population Genetics* (pp. 113–128).

461   https://doi.org/10.1093/oso/9780198838944.003.0006

462   Dapper, A. L., & Payseur, B. A. (2017). Connecting theory and data to understand

463   recombination rate evolution. *Philosophical Transactions of the Royal Society of*

464   *London. Series B, Biological Sciences*, *372*(1736), 20160469.

465   Dapper, A. L., & Payseur, B. A. (2018). Effects of Demographic History on the Detection

466   of Recombination Hotspots from Linkage Disequilibrium. *Molecular Biology and*

467   *Evolution*, *35*(2), 335–353.

468   DeWitt, W. S., Harris, K. D., Ragsdale, A. P., & Harris, K. (2021). Nonparametric

469   coalescent inference of mutation spectrum history and demography. *Proceedings of*

470   *the National Academy of Sciences of the United States of America*, *118*(21).

471   https://doi.org/10.1073/pnas.2013798118

472   Dias-Alves, T., Mairal, J., & Blum, M. G. B. (2018). Loter: A Software Package to Infer

473   Local Ancestry for a Wide Range of Species. *Molecular Biology and Evolution*, *35*(9),

474   2318–2326.

475   Dréau, A., Venu, V., Avdievich, E., Gaspar, L., & Jones, F. C. (2019). Genome-wide

476   recombination map construction from single individuals using linked-read

477   sequencing. *Nature Communications*, *10*(1), 4309.

478   Dumont, B. L., & Payseur, B. A. (2008). Evolution of the genomic rate of recombination

479   in mammals. *Evolution; International Journal of Organic Evolution*, *62*(2), 276–294.

480   Ellstrand, N. C., & Rieseberg, L. H. (2016). When gene flow really matters: gene flow in

481   applied evolutionary biology. *Evolutionary Applications*, *9*(7), 833–836.

482   Haddrill, P. R., Charlesworth, B., Halligan, D. L., & Campos, J. L. (2014). The relation

483   between recombination rate and patterns of molecular evolution and variation in

484   Drosophila melanogaster. *Molecular Biology and Evolution*, *31*(4), 1010–1028.

485   Haller, B. C., & Messer, P. W. (2019). SLiM 3: Forward Genetic Simulations Beyond the

486   Wright-Fisher Model. *Molecular Biology and Evolution*, *36*(3), 632–637.

487   Hermann, P., Heissl, A., Tiemann-Boege, I., & Futschik, A. (2019). LDJump: Estimating
488   variable recombination rates from population genetic data. *Molecular Ecology*
489   *Resources*, *19*(3), 623–638.

490   Hunter, C. M., Huang, W., Mackay, T. F. C., & Singh, N. D. (2016). The Genetic
491   Architecture of Natural Variation in Recombination Rate in Drosophila
492   melanogaster. *PLoS Genetics*, *12*(4), e1005951.

493   Johnston, S. E., Bérénos, C., Slate, J., & Pemberton, J. M. (2016). Conserved Genetic
494   Architecture Underlying Individual Recombination Rate Variation in a Wild
495   Population of Soay Sheep (Ovis aries). *Genetics*, *203*(1), 583–598.

496   Kamm, J. A., Spence, J. P., Chan, J., & Song, Y. S. (2016). Two-Locus Likelihoods Under
497   Variable Population Size and Fine-Scale Recombination Rate Estimation. *Genetics*,
498   *203*(3), 1381–1399.

499   Korunes, K. L., Samuk, K., & Noor, M. A. F. (2021). Disentangling Types of Linked
500   Selection Using Patterns of Nucleotide Variation in Drosophila pseudoobscura. In
501   *Population Genomics* (pp. 1–22). Springer International Publishing.

502   Li, H., & Stephan, W. (2006). Inferring the demographic history and rate of adaptive
503   substitution in Drosophila. *PLoS Genetics*, *2*(10), e166.

504   Linck, E., & Battey, C. J. (2019). *On the relative ease of speciation with periodic gene flow*.
505   https://doi.org/10.1101/758664

506   Lloyd, A., Morgan, C., H Franklin, F. C., & Bomblies, K. (2018). Plasticity of Meiotic
507   Recombination Rates in Response to Temperature in Arabidopsis. *Genetics*, *208*(4),
508   1409–1420.

509   Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology &*

510   *Evolution*, *20*(5), 229–237.

511   McVean, G. (2007). Linkage Disequilibrium, Recombination and Selection. In D. J.

512   Balding, M. Bishop, & C. Cannings (Eds.), *Handbook of Statistical Genetics* (pp. 909–

513   944). John Wiley & Sons, Ltd.

514   McVean, G., & Auton, A. (2007). LDhat 2.1: a package for the population genetic analysis

515   of recombination. *Department of Statistics, Oxford, OX1 3TG, UK*.

516   http://www.stats.ox.ac.uk/~mcvean/LDhat/manual.pdf

517   Morjan, C. L., & Rieseberg, L. H. (2004). How species evolve collectively: implications of

518   gene flow and selection for the spread of advantageous alleles. *Molecular Ecology*,

519   *13*(6), 1341–1356.

520   Nei, M., & Li, W.-H. (1973). LINKAGE DISEQUILIBRIUM IN SUBDIVIDED

521   POPULATIONS. In *Genetics* (Vol. 75, Issue 1, pp. 213–219).

522   https://doi.org/10.1093/genetics/75.1.213

523   Ochoa, A., & Storey, J. D. (2019). New kinship and FST estimates reveal higher levels of

524   differentiation in the global human population. *BioRxiv*.

525   https://www.biorxiv.org/content/10.1101/653279v1.abstract

526   Ohta, T. (1982). Linkage disequilibrium with the island model. *Genetics*, *101*(1), 139–155.

527   Peñalba, J. V., & Wolf, J. B. W. (2020). From molecules to populations: appreciating and

528   estimating recombination rate variation. *Nature Reviews. Genetics*, *21*(8), 476–492.

529   Peterson, A. L., Miller, N. D., & Payseur, B. A. (2019). Conservation of the genome-wide

530   recombination rate in white-footed mice. *Heredity*, *123*(4), 442–457.

531   Peterson, A. L., & Payseur, B. A. (2021). Sex-specific variation in the genome-wide

532   recombination rate. *Genetics*, *217*(1), 1–11.

26

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

535   Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M. A.
536   F., Mehlig, B., & Westram, A. M. (2017). Interpreting the genomic landscape of
537   speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary*
538   *Biology*, *30*(8), 1450–1477.

539   R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation
540   for Statistical Computing, Vienna, Austria. https://www.R-project.org/

541   Ritz, K. R., Noor, M. A. F., & Singh, N. D. (2017). Variation in Recombination Rate:
542   Adaptive or Not? *Trends in Genetics: TIG*, *33*(5), 364–374.

543   Rommel Fuentes, R., Hesselink, T., Nieuwenhuis, R., Bakker, L., Schijlen, E., van
544   Dooijeweert, W., Diaz Trivino, S., de Haan, J. R., Sanchez Perez, G., Zhang, X.,
545   Fransz, P., de Jong, H., van Dijk, A. D. J., de Ridder, D., & Peters, S. A. (2019).
546   Meiotic recombination profiling of interspecific hybrid F1 tomato pollen by linked
547   read sequencing. *The Plant Journal: For Cell and Molecular Biology*.
548   https://doi.org/10.1111/tpj.14640

549   Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., & Bierne, N. (2016).
550   Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic
551   Divergence. *PLoS Biology*, *14*(12), e2000234.

552   Samuk, K., Manzano-Winkler, B., Ritz, K. R., & Noor, M. A. F. (2020). Natural Selection
553   Shapes Variation in Genome-wide Recombination Rate in Drosophila
554   pseudoobscura. *Current Biology: CB*, *30*(8), 1517–1528.e6.

555   Samuk, K., Owens, G. L., Delmore, K. E., Miller, S. E., Rennison, D. J., & Schluter, D.
556   (2017). Gene flow and selection interact to promote adaptive divergence in regions

557   of low recombination. *Molecular Ecology*, *26*(17), 4378–4390.

558 Schumer, M., Xu, C., Powell, D. L., Durvasula, A., Skov, L., Holland, C., Blazier, J. C.,

559   Sankararaman, S., Andolfatto, P., Rosenthal, G. G., & Przeworski, M. (2018). Natural

560   selection interacts with recombination to shape the evolution of hybrid genomes.

561   *Science*, *360*(6389), 656–660.

562 Slarkin, M. (1985). Gene Flow in Natural Populations. *Annual Review of Ecology and*

563   *Systematics*, *16*(1), 393–430.

564 Smukowski Heil, C. S., Ellison, C., Dubin, M., & Noor, M. A. F. (2015). Recombining

565   without Hotspots: A Comprehensive Evolutionary Portrait of Recombination in

566   Two Closely Related Species of Drosophila. *Genome Biology and Evolution*, *7*(10),

567   2829–2842.

568 Spence, J. P., & Song, Y. S. (2019). Inference and analysis of population-specific fine-

569   scale recombination maps across 26 diverse human populations. *Science Advances*,

570   *5*(10), eaaw9206.

571 Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., & Smadja, C. M. (2017).

572   Variation in recombination frequency and distribution across eukaryotes: patterns

573   and processes. *Philosophical Transactions of the Royal Society of London. Series B,*

574   *Biological Sciences*, *372*(1736). https://doi.org/10.1098/rstb.2016.0455

575 Stumpf, M. P. H., & McVean, G. A. T. (2003). Estimating recombination rates from

576   population-genetic data. *Nature Reviews. Genetics*, *4*(12), 959–968.

577 Suvorov, A., Kim, B. Y., Wang, J., Armstrong, E. E., Peede, D., D'Agostino, E. R. R.,

578   Price, D. K., Wadell, P., Lang, M., Courtier-Orgogozo, V., David, J. R., Petrov, D.,

579   Matute, D. R., Schrider, D. R., & Comeault, A. A. (2021). Widespread introgression

580   across a phylogeny of 155 Drosophila genomes. In *bioRxiv* (p. 2020.12.14.422758).

581   https://doi.org/10.1101/2020.12.14.422758

582 Waples, R. S., & Gaggiotti, O. (2006). What is a population? An empirical evaluation of

583   some genetic methods for identifying the number of gene pools and their degree of

584   connectivity. *Molecular Ecology*, *15*(6), 1419–1439.

585 Wickham, H. (2017). The tidyverse. *R Package Ver. 1.1*.

586   https://slides.nyhackr.org/presentations/The-Tidyverse_Hadley-Wickham.pdf

587 Xu, P., Kennell, T., Gao, M., Human Genome Structural Variation Consortium,

588   Kimberly, R. P., & Chong, Z. (2020). MRLR: unraveling high-resolution meiotic

589   recombination by linked reads. *Bioinformatics* , *36*(1), 10–16.

590

Supplemental Material
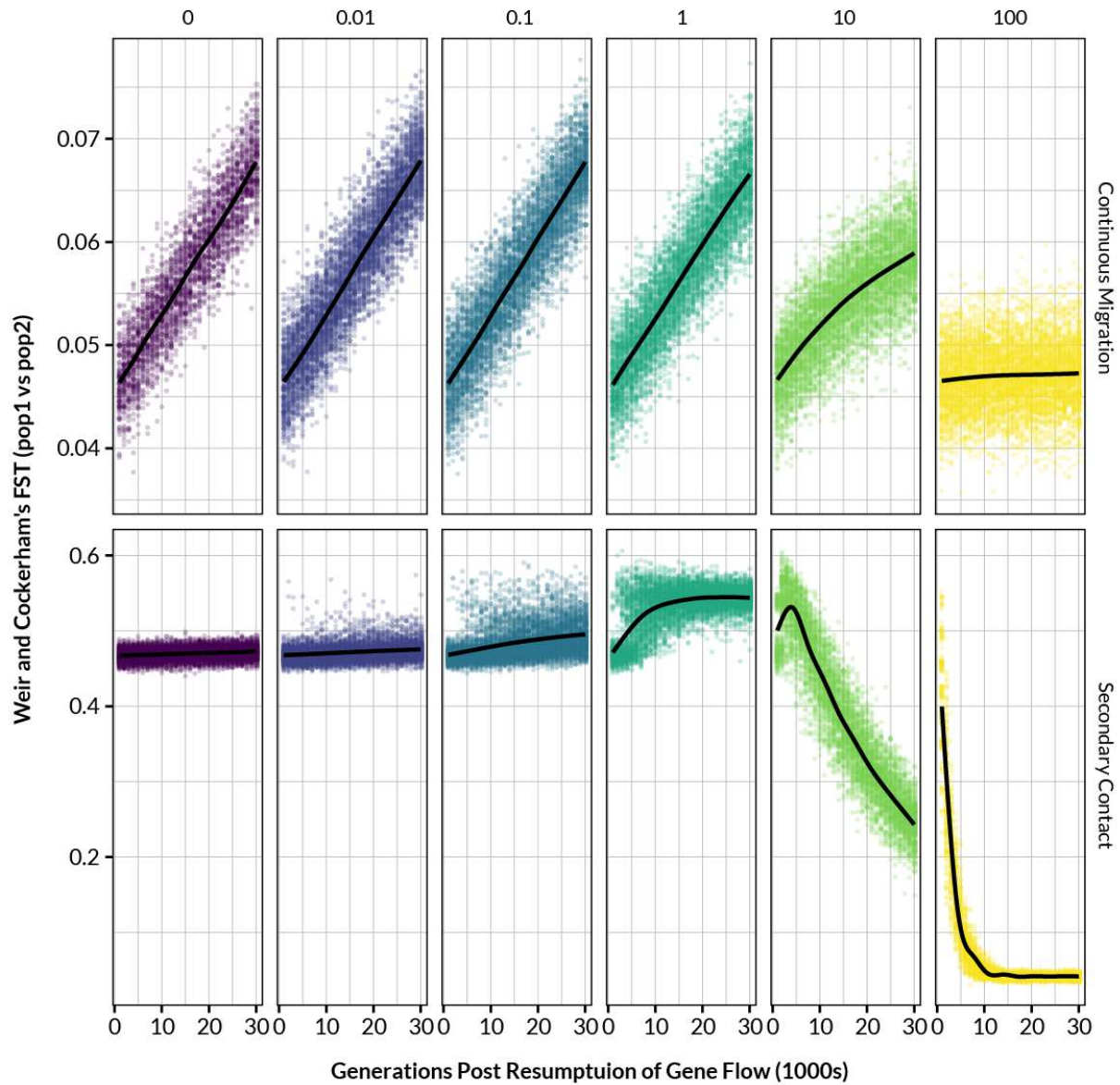
Figure S1 | Weir and Cockerham's FST between simulated populations as a function of time in generations under various combinations of migration rate (columns, N_em) and isolation scenario (rows). Black lines are smoothed LOESS fits. Note the difference in y-axis scales between the rows.