# Gene flow biases population genetic inference of recombination rate

K Samuk, MAF Noor

## Introduction

Recombination rate is a key mediator of population genetic and evolutionary processes. For example, local rates of recombination are the chief determinant of patterns of genetic diversity throughout the genome via the effects of linked selection (cite Cutter review). Genome-wide rates of recombination also modulate diverse processes such as adaptation, speciation, and introgression (cite Stapley et al, cite Dapper & Payseur). There is also a growing appreciation that recombination rate is itself a trait that varies and evolves (cite Stapley et al, cite Ritz et al). Accordingly, there has been greatly increased interest in efficient and accurate methods for estimating recombination rate.

Currently, there are a variety of methods for estimating recombination rate, each with their own strengths and weaknesses. These fall into two broad classes of methods: direct and indirect (cite Penalba review). 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 in the context of a genetic cross (cite LepMap paper). While the resolution of this approach is limited only by marker density and the size of the cross, larger crosses are highly laborious to carry out. Further, identifying suitable diagnostic mapping markers can be limiting in some cases (e.g. in a highly homozygous inbred line). Direct sequencing of pools of 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 (cite Dreau et al. 2019). Cytological methods bypass this requirement by visualizing recombination-associated protein complexes in cell populations undergoing meiosis (cite Peterson et al. 2019). 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).

All direct methods of measuring recombination rate are somewhat laborious. As such, with advent of population-scale genome sequencing, there has been increased interest in indirect measures of recombination rate using approaches using population genetics. Chief among these are model-based methods that infer rates of recombination rate from patterns of linkage disequilibrium (LD), (cite LDhat, LDhelmet, pyrho). First introduced by McVean & Auton (cite), these methods attempt to estimate recombination rates by statistically fitting recombination rates (derived from population genetic models) to observed patterns of LD. Rather than inferring crossover rate directly, LD-based estimators infer a *population scaled recombination rate,* ⍴ = 4Nec, where Ne is the effective population size and c is the theoretical per-generation recombination rate. LD-based methods are highly attractive because they (1) generally only require population-scale genomic data and (2) the process of estimation is fast, often only requiring several computational hours or less (cite Spence pyrho paper). Accordingly, LD-based estimates of recombination rate have become extremely popular, and now vastly outnumber direct measures in the literature (cite Penalba review, cite Stapley review).

LD-based estimators of recombination rate generally assume that samples are drawn from populations that are neutrally-evolving and have reached population genetic equilibrium in a number of ways. First, these methods assume that the populations being studied have reached (or nearly reached) an “LD equilibrium” between recombination and population scaled mutation, such that LD accurately reflects patterns of recombination rate. Secondly, the assumption of neutrality assumes that any form of selection that might distort patterns of LD (e.g. sweeps), have not recently occured. Finally, these methods make a general assumption that demographic processes that distort genome-wide patterns of LD, such as population size changes or migration, have not occured (recall that ⍴ is direct dependant on Ne). Violations of these equilibrium assumptions can result in misestimation of recombination rate. For example, Dapper et al. showed that recombination estimates from LDhat are highly sensitive to changes in population size (addressed by Spence et al. using a new method that attempts to account for this effect).

Along with changes in population size, another process that can cause non-equilibrium conditions in natural populations is gene flow. Gene flow from other populations can have complex effects on patterns of LD in a population, ranging from large and variable increases due to allele frequency differences between populations, to decreases in LD as populations become coupled and increase local Ne (cite Nei 1974). While it is now widely accepted that gene flow is ubiquitous in natural populations (cite Mallet 2005, cite Matute 2021 review), and there has not been a systematic study of the effects of gene flow on LD-based measures of recombination. Further, it remains unclear how gene flow (or any other violation of assumptions) impacts our ability to detect differences in recombination rate between populations using LD-based methods.

Here, we address these issues using modern forward-time population genetic simulations. We attempt to answer two specific questions. First, how does gene flow between populations affect the precision and accuracy of LD-based estimates of recombination rate? 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 that complicate LD-based inference of recombination rate, and not perform an exhaustive exploration of parameter space. Overall we hope to ground expectations about the utility of LD-based estimators of recombination rate in natural plant and animal populations, and highlight areas of future development.

## Methods

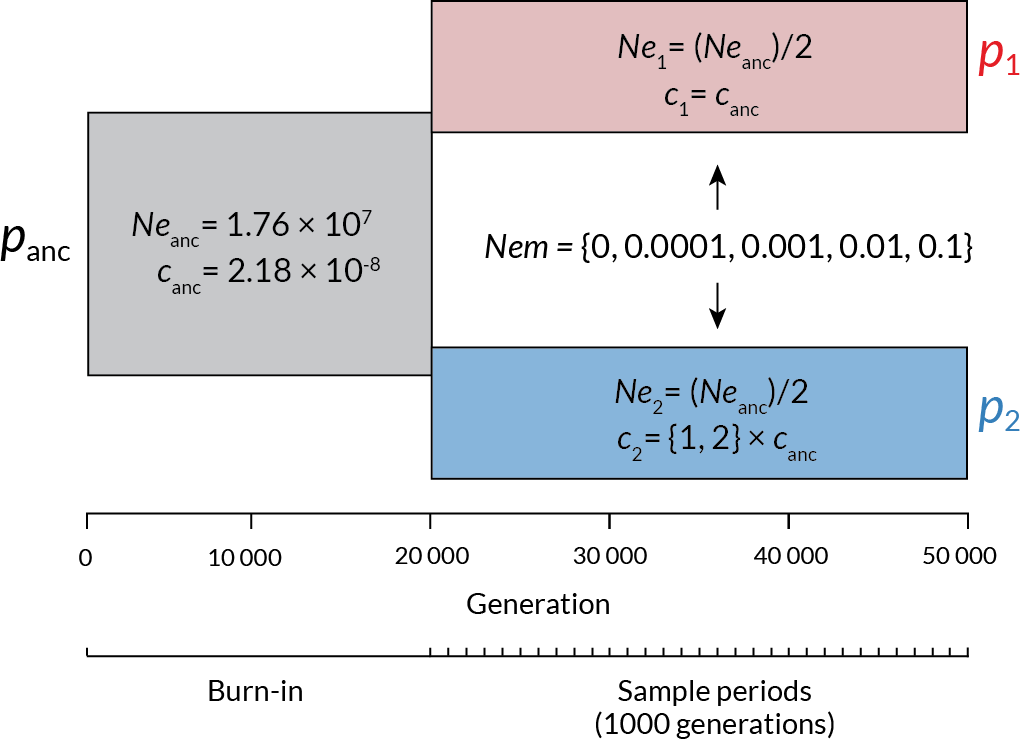
#### Code availability

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

#### Forward time simulations with SLiM

We simulated a scenario in which a neutrally-evolving population at demographic equilibrium diverges into two sub-populations, varying the level of (1) gene flow and (2) divergence in genome-wide recombination rate (Figure 1). All simulations were performed using SLiM 3.3 (cite slim). Each individual was diploid, with a genome composed of 100kb of neutral loci. We used estimates of effective population size, mutation rate, and (empirical) recombination rate from natural populations of *Drosophila melanogaster* (cite popsim, Mutation = X, Recombination = 2.X, average of chromosome 2R, Ne = 1.6M, cite cite). Recombination and mutation rates were modelled as uniform across the 100kb genome. Following standard practice for forward time simulations, all simulations were run with a *in silico* population size of N=1000, and mutation and recombination rates scaled by a factor of Ne/N as per the SLiM manual.

To explore the effects of gene flow, we varied the amount of gene flow over five orders of magnitude: 0, 0.0001, 0.001, 0.01, 0.1, in units of the product of the effective population size and the migration rate (Nem). To explore how gene flow impacts the detection of population differences in recombination rate, we also modelled scenarios where recombination remains constant in both subpopulations, or instantaneously increases by a factor of 2 in one of the two subpopulations. Note this is a somewhat extreme case chosen for illustrative purposes, and the majority of known examples of divergence in recombination rate are much smaller (e.g. a factor of 1.1 in the case of Samuk et al.).



**Figure 1** | The structure of the forward-time simulations performed in SLiM. Time in generations is shown along the x-axis, and the populations in existence at a given time are shown as rectangles. Panc = the ancestral population, P1 = the daughter population with unchanged recombination rate, and P2 = the daughter population with increased recombination rate. Effective population sizes (Ne, individuals) and recombination rates (cM/Mb) are shown for each population, with the values for the daughter populations shown in relation to the value in the ancestral population. Variable elements of the simulation are shown in braces.

Each simulation began with a single population of size Ne, which evolved for a 20 000 generation burn-in period (following the general practice of a 10-20Ne burn-in period). This initial period was followed by divergence at generation 20 000 into two subpopulations, each with size Ne/2. Gene flow (if applicable) began at t=20 000 and was continuous and bidirectional. Changes in recombination rate also occurred at t=20 000, and applied to all individuals in subpopulation two only. At t=20 000, and thereafter in intervals of 1000 generations, we collected a random sample of 25 individuals from each population and saved their genotypic states in VCF format. We stopped the simulations at t=50 000 (For a total of 50 000 generations: 20 000 burn in and 30 000 following divergence). Each parameter combination was replicated 100 times, for a total number of ~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 (cite Spence) for estimation in this study. It shares its statistical foundation with the most widely used LD-based estimators (LD-hat, LD-helmet), while also having the ability to account for changes in effective population size (such as we are modelling here).

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 a lookup table that accounts for a change of Ne/2 at 20000 generations in the past (“--popsizes 860300, 1720600”, “--epochtimes 20000”) . We used the built-in methods to infer the hyperparameters of window size (best fit 100) and block penalty (best fit 1000). Using this baseline, we inferred recombination rate using the VCF data from both subpopulations at each time point, for a total of ~96 000 pyrho fits. All computation was performed using the Duke University Computing Cluster, running CentOS Version 8.

#### Statistical analyses

We performed all data preparation, exploratory analyses, and plots using the tools of the tidyverse package in R 4.0.3 (cite R, cite tidyverse). The scripts for the statistical analysis are available at http://github.com/ksamuk/LD\_gene\_flow\_recomb.

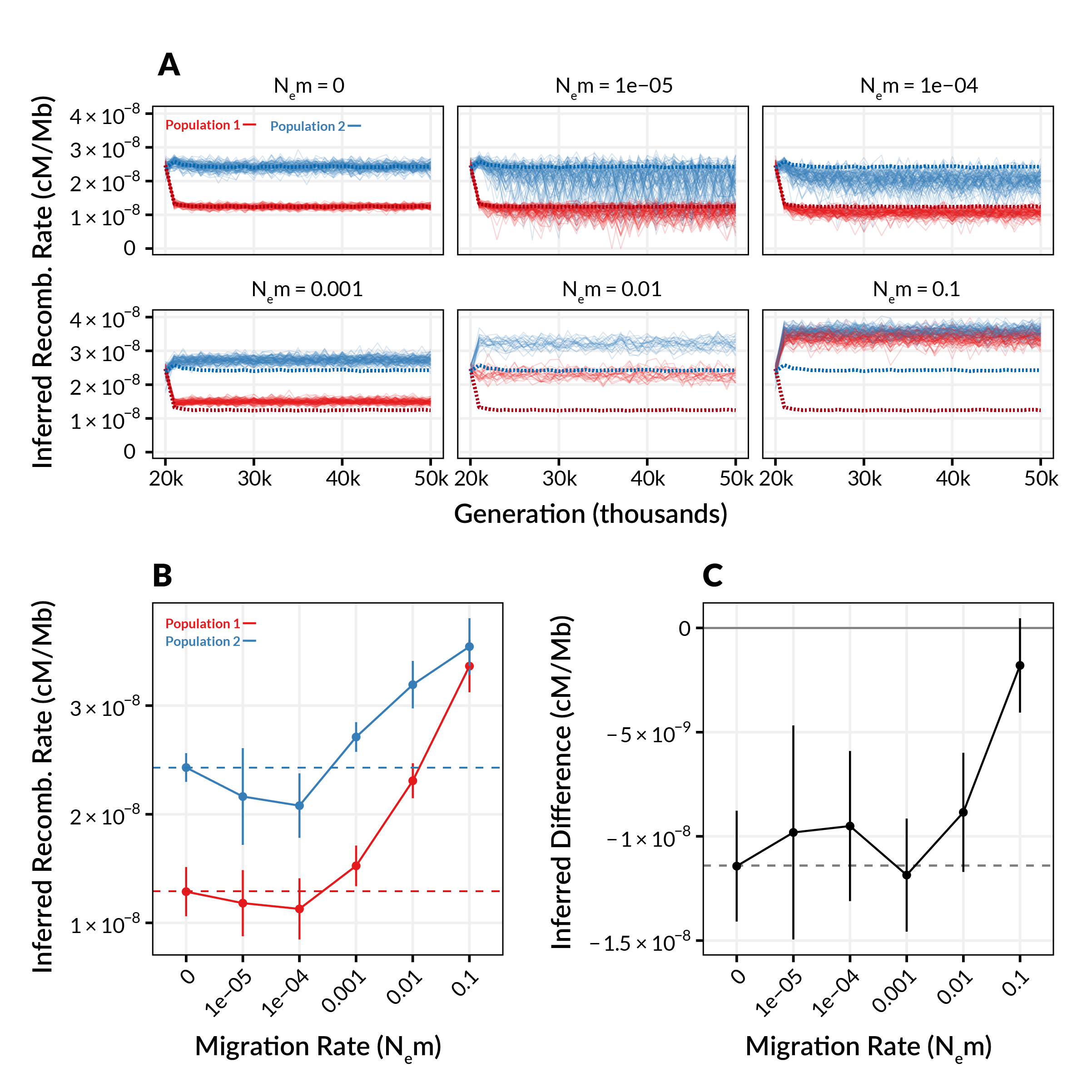
To examine how gene flow between populations affects the accuracy of LD-based estimate s of recombination, and the context of the various factors explored in our simulations, we performed an analysis of variance using a linear mixed model with gaussian errors fitted in R (cite version) via the lmer() function from the lme4 package (cite Bates). This model had the following form: Recombination rate = (1|replicate) + gene flow + recombination change + gene flow:recombination change + error, where (1|[factor]) denotes a random effect and “:” denotes an interaction.

## Results

Our simulations suggest that gene flow between populations biases LD-based estimates of recombination rate within populations. When the migration rate is low (NeM ~ 0.0001, Figure 2A), LD is elevated, which translates to a statistically significant downward bias in estimates of recombination rate (Type III Analysis of Deviance, p < 2x10-22, see Table S1, more stats TBD). In contrast, when migration rate is high (NeM > 0.001), the subpopulations become coupled, effectively behaving as a single large population with a larger Ne than any single subpopulation. This results in a *decrease* in LD, and significantly upwardly biased estimated values of recombination rate (Figure 2A, see NeM = 0.01 and NeM = 0.1, Figure 2B; more stats TBD). In addition to the biases discussed above, gene flow (at any level) resulted in highly elevated variance in recombination rate estimates (Figure 2B, more stats TBD). The exception to this appeared at NeM = 0.001, which appears to be a fortuitous inflection point between the downward and upward biases. This pattern (increased variance and increasing bias) was also observable in simulations where recombination rate diverged between populations (Figure 2A), and in simulations where it remained constant (Figure S1).

The decrease in precision and accuracy due to gene flow also affects the ability to observe *differences* in recombination rate between populations. As in the single population case, low levels of gene flow result in increased variance in estimates of recombination rate differences between populations (Figure 2C). Higher levels of gene flow result in the aforementioned coupling effect, which while elevating observed recombination rates within populations, results in a decrease in the magnitude of the observed difference in recombination rate between populations.

The magnitude of the increased variance and coupling effects discussed depended on both the migration rate per se and the time since divergence. For higher levels of gene flow, equilibrium levels of LD occurred around generation 10 000 (Figure 2A, NeM 0.001 and above). The increased variance in LD at low levels of migration was weak initially, but quickly increased by generation ~10 000, and reach edits maximum variance around generation 30 000, never apparently decreasing.



**Figure 2** | The relationship between inferred recombination rate and the migration rate in simulated populations where recombination rate diverges between populations. (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 Nem = 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 Nem = 0). (C) The inferred *difference* in recombination rate between population 1 and population 2 (Pop1 - Pop2) as a function of migration rate. Points and errors bars are as in B.

## Discussion

Accurate estimates of recombination rate are key to understanding the causes and consequences of recombination rate variation in natural populations. With the advent of genome-wide sequencing data, LD-based estimators of recombination rate have become widely used in a large variety of taxa. 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, use simulations to show that gene flow can affect both the precision and accuracy of LD-based estimators of recombination rate, depending on the level of gene flow and the direction of recombination divergence.

The general result we report here, i.e. gene flow distorts LD, will not be surprising to any student of population genetics. Nei (1974) very clearly outlined how gene flow between populations can affect LD, and described both the increase in LD magnitude and variance at low migration rates, the coupling effect at higher rates. Our study shows how these predictions play out with real world methods and data, and also provides a sense of the magnitude of the potential degree of misestimation. For comparison, our recent study of population-level differences in recombination rate in Drosophila pseudoobscura revealed interpopulation differences on the order of 10% (measured using replicated linkage maps in each population). Using LD-based estimators, even in the absence of gene flow, such a difference could be readily observed by chance (if gene flow is low), or missed altogether due to coupling (if gene flow is high). This should give pause to anyone planning on using LD-based methods to infer recombination rate in non-equilibrium populations.

One key question is whether there are methods to control for or counteract the increased variance and/or biases caused by gene flow. One method to attempt to correct for the biases and increased variance at low levels of gene flow could be to identify and remove introgressed haplotypes from datasets prior to inferring recombination rate. This would require samples from the source populations, such that population of origin could be assigned to haplotype blocks (e.g. cite IBIS). However, this method would only work if gene flow is infrequent 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 between populations at higher levels of gene flow will likely be much more difficult (perhaps impossible) to counteract. One approach may be to attempt to jointly estimate an isolation-with-migration model along with population-specific recombination rates (as has been done with mutation rates, cite Harris). However, given the existing complexity and uncertainty in inferring demographic models, we suspect it may be even more difficult to disentangle the complex interdependencies between gene flow, population size, and estimates of recombination rate.

Together with previous work (cite Dapper et al), our results suggest that LD-based estimators of recombination rate may have limited utility for studying variation in recombination rates in many natural populations. LD-base estimators are likely only appropriate when populations can be assumed to be evolving in the absence of any gene flow from a diverged population, and reached demographic equilibrium. However, if this is true, why do published LD-based estimates of recombination rate correlate well with direct estimates, e.g. from genetic maps? ( cite LDhat paper, cite Smukowski-Heil et al. 2015). There are several considerations. First, the correlations that have been reported are high but imperfect (~R^2 of 0.7). Second, simple correlations between LD-based and empirical estimates cannot detect genome-wide differences in recombination rate, such as those due to the coupling effects we observed. Such effects would be visible as differences in theintercept of a linear regression, for example. Finally, the species where these correlations have been examined (humans and Drosophila) are not known to be subject to substantial gene flow from *divergent* populations and may meet demographic assumptions readily (cite human FST paper, cite drosophila gene flow paper) . While such assumptions may be reasonable for human 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 diverged in the presence of gene flow, which have lately been the subject of increased research interest (cite speciation with gene flow review).

While we only focused on a single implementation of one type of LD-base estimator of recombination (pyrho), it is likely that all other population genetic methods will also suffer from effects we describe here. LD is the “information” used by all estimators, either directly (as in methods like LDjump) or indirectly (as in machine learning methods like ReLERN, cite Adrion). 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 align with reality are readily violated by gene flow.

For the purposes of simplicity, our study focused on the inference of a 100kb segment with a uniform recombination rate. However, most estimators of recombination rate are explicitly designed to infer *variable* recombination rates, usually in genomic windows of some size (cite LDhat). A key future extension of the work presented here would be to examine how gene flow impacts inference of the recombination landscape. Rather than being uniformly increased or decreased in the presence of gene flow, we expect that inference of the recombination landscape will be highly sensitive to the state of breakdown of introgressed haplotypes, and the extent at which these happen to exist in the individuals sampled for study (cite haplotype decay paper). For example, regions with recently introgressed haplotypes will appear to have “lower” recombination than regions with older introgression, even in the absence of any actual differences in recombination rate.

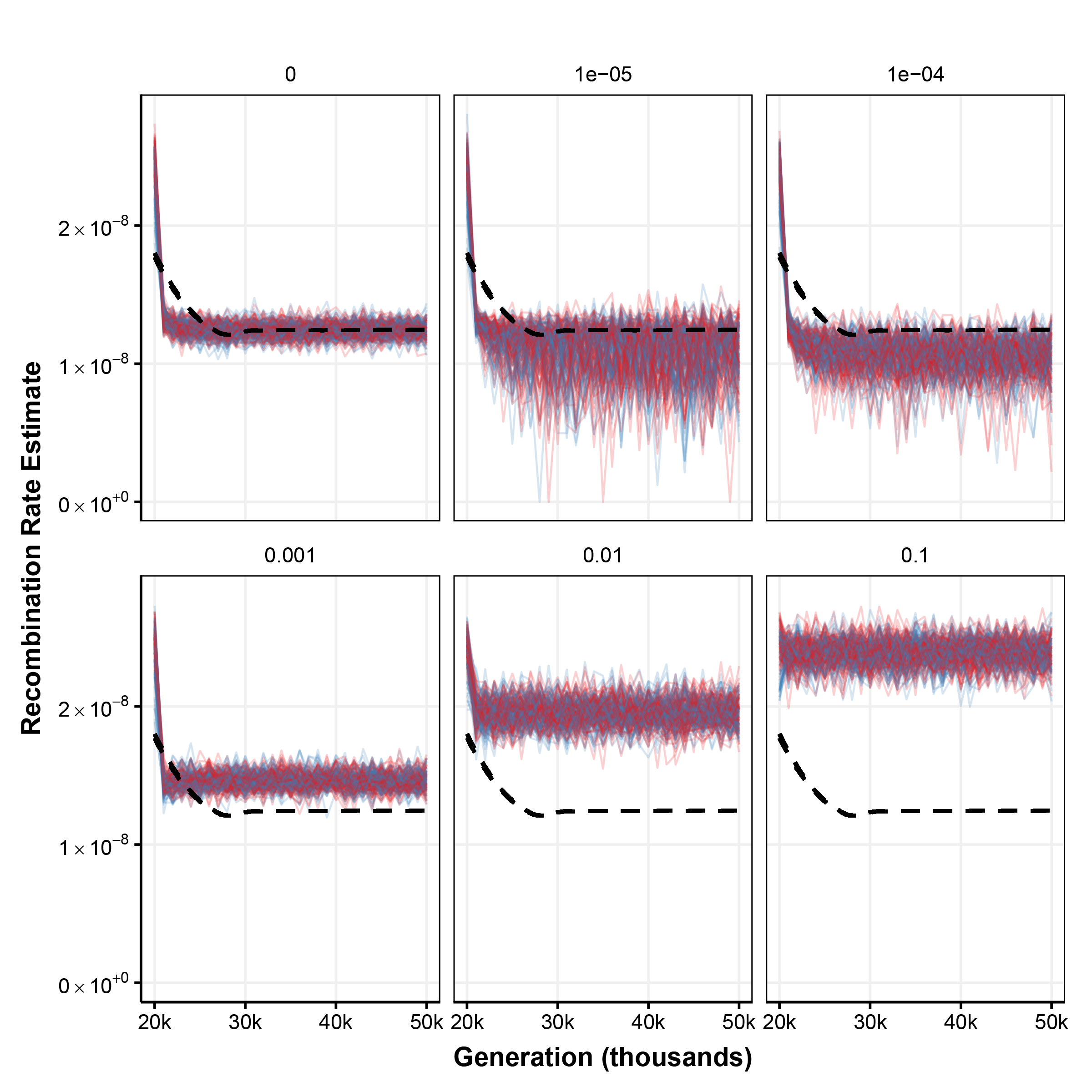
#### Conclusion

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 large biases and decreases in precision. 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 measures of recombination rate over population genetic estimators. With the increased ease and low cost of creating traditional linkage maps or performing gamete sequencing, investigators interested in studying variation in recombination rate should strongly prefer these methods over LD-based approaches.

## References

[TBD]

## Supplemental Material



**Figure S1** | The relationship between inferred recombination rate and the migration rate in simulated populations where recombination rate *remains the same* between diverging populations. Each plot shows inferred recombination rates for simulation replicates (transparent lines) of population 1 (red, unchanged recombination) and population 2 (blue, also unchanged recombination) for a single migration rate. The black dashed line shows the expected inferred value in the absence of gene flow (inferred from Nem = 0).

**Table S1** | Scaled regression coefficients for the linear mixed model of the relationship between migration rate, recombination rate divergence, and inferred recombination rate. Simulation replicate and generation were both modeled as random effects.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Inferred Recomb. Rate** | | |
| *Predictors* | *Estimates* | *CI* | *p* |
| (Intercept) | 1.76 | 1.63 – 1.89 | **<0.001** |
| Migration Rate | 5.41 | 5.37 – 5.45 | **<0.001** |
| Recomb. Divergence | 0.03 | 0.03 – 0.03 | **<0.001** |
| Mig. Rate x Recomb. Div. | 4.91 | 4.85 – 4.97 | **<0.001** |
| **Random Effects** | | | |
| σ2 | 0.10 | | |
| τ00 replicate | 0.02 | | |
| τ00 generation | 0.22 | | |
| ICC | 0.70 | | |
| N generation | 51 | | |
| N replicates | 4171 | | |