



# Estimating recombination using only the allele frequency spectrum

Matthew W. Hahn <sup>1,2,\*</sup> Sarthak R. Mishra <sup>2</sup>

<sup>1</sup>Department of Biology, Indiana University Bloomington, 1001 E. 3rd St., Bloomington, IN 47405, USA

<sup>2</sup>Department of Computer Science, Indiana University Bloomington, 700 N. Woodlawn Ave., Bloomington, IN 47405, USA

\*Corresponding author: Department of Biology, Indiana University Bloomington, 1001 E. 3rd St., Bloomington, IN 47405, USA. Email: mwh@iu.edu

Standard methods for estimating the population recombination parameter,  $\rho$ , are dependent on sampling individual genotypes and calculating various types of disequilibria. However, recent machine learning (ML) approaches to estimating recombination have used pooled sequencing data, which does not sample individual genotypes and cannot be used to calculate disequilibria beyond the length of a single sequence read. Motivated by these results, this study examines the “black box” of such ML methods to understand what signals are being used to infer recombination rates. We find that it is indeed possible to estimate recombination solely using the allele frequency spectrum, and we provide a genealogical interpretation of these results. We further show that even a simplified representation of the allele frequency spectrum can be used to estimate recombination. We demonstrate the accuracy of such inferences using both simulations and data from humans. These results offer a new way to understand the effects of recombination on patterns of sequence data, as well as providing an example of how the internal workings of ML methods can give insight into biological processes.

**Keywords:** coalescent; linkage disequilibrium; genealogies; machine learning

## Introduction

Recombination is a fundamental biological process that plays an important role in evolution (Johnston 2024). While crosses between individuals and the genotyping of a large number of offspring are often used to infer the meiotic recombination rate,  $c$ , the population recombination parameter,  $\rho$  ( $=4N_e c$ ), can be inferred from a small sample of unrelated individuals. The magnitude of this parameter reflects the history of recombination in the sample across many thousands of generations but is often strongly correlated with the underlying meiotic recombination rate (e.g. McVean et al. 2004; Stevison et al. 2016).

There are multiple common ways to estimate  $\rho$  (reviewed in Hahn 2018, chapter 4; Peñalba and Wolf 2020). Possibly, the most widely used set of methods are based on gametic linkage disequilibrium (LD), using individually phased haplotypes to estimate the association between alleles on chromosomes. Measures of gametic LD can then be used to estimate  $\rho$  (Sved 1971; Weir and Hill 1986; McVean 2002), or haplotypes can be used directly (e.g. Hudson 1987; Wakeley 1997; Wall 2000). If phased haplotypes are not available, another form of LD can still be calculated from diploid genotypes: genotypic LD (Weir 1979). The very popular (and accurate) class of methods that estimate  $\rho$  using composite likelihood (Hudson 2001; McVean et al. 2002; Chan et al. 2012; Kamm et al. 2016; Spence and Song 2019) can all use either phased haplotypes (i.e. gametic LD) or unphased genotypes (i.e. genotypic LD). Finally, a newer set of approaches based solely on whether positions are heterozygous or homozygous—without respect to the particular alleles or genotypes at a site—have been used to calculate so-called zygotic LD and consequently  $\rho$  (Haubold et al. 2010;

Barroso et al. 2019; Setter et al. 2022). Despite the relative lack of resolution in the recombination rate using zygotic LD, such approaches are also highly accurate (Dutheil 2024).

In the past few years, machine learning (ML) methods have become a useful and accurate approach for multiple types of inference in population genetics (Schridder and Kern 2018; Korfmann et al. 2023; Huang et al. 2024). Machine learning methods are especially useful in dealing with messy data: in the case of estimating  $\rho$ , this might mean incorrectly inferred haplotypes or genotypes. Indeed, multiple ML approaches for estimating  $\rho$  have been introduced over the past dozen years (Lin et al. 2013; Gao et al. 2016; Flagel et al. 2019; Hermann et al. 2019), including ReLERNN (Recombination Landscape Estimation using Recurrent Neural Networks), a deep learning tool that can accurately infer  $\rho$  from suboptimal data (Adrion et al. 2020).

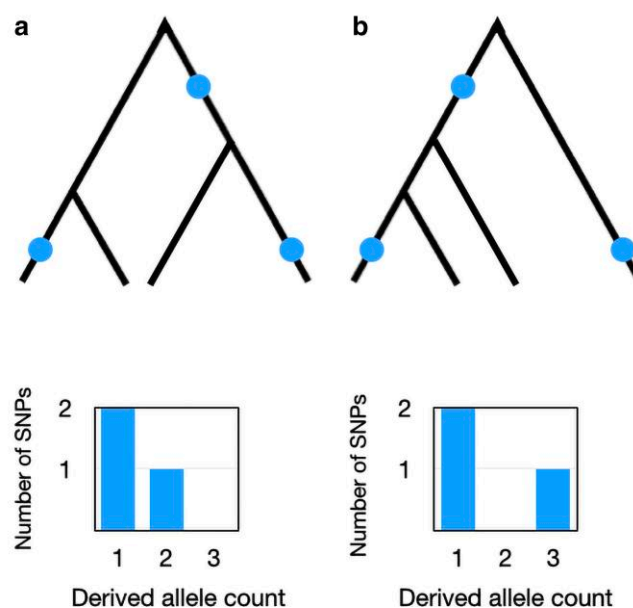
Most interestingly, ReLERNN is also able to accurately infer  $\rho$  from pooled sequencing data. Pooled sequencing (sometimes called “pool-seq”; Schlötterer et al. 2014) provides only allele frequencies at each genomic position, as no barcodes or labels are associated with each sampled individual in the pool. While there have been previous methods that could infer very short-range LD from pooled sequencing (Feder et al. 2012), these rely on SNPs found in the same read and are therefore limited to short distances. In contrast, ReLERNN does not take any information about sequence reads into account—the input contains only a list of SNP positions and allele frequencies within a genomic window. Although no obvious type of disequilibrium can be calculated from such data, Adrion et al. (2020) showed that ReLERNN can very accurately infer  $\rho$  across larger distances.

Machine learning methods can learn from messy, high-dimensional data, but are also prone to picking up on unintended signals provided by, for instance, the order in which data are presented, seemingly innocuous data labels, or other non-meaningful aspects of the training set (Bernett et al. 2024). Putting aside the possibility of such data leakage, there are multiple signals associated with recombination that ReLERNN could be used for inferences from pooled sequencing (Adrión and colleagues do not speculate as to the source of the signal). First, the input to ReLERNN implicitly encodes the number of SNPs in a window as the number of columns in the dataset. As there is a near-universal correlation in natural populations between the number of SNPs in a region—often represented by the population mutation parameter,  $\theta$  ( $=4N_e\mu$ )—and the population recombination parameter (Cutter and Payseur 2013), it is possible that ReLERNN could use this relationship to estimate  $\rho$ . However, a strong relationship between  $\theta$  and  $\rho$  can only arise in non-neutral scenarios, and Adrión et al. (2020) show that their method is still accurate in neutral, equilibrium populations. Second, the input to ReLERNN contains the genomic position of each SNP in a window. While it is not obvious what sort of information the distance between variable positions might contain about recombination, it is possible that it is using this information. Finally, and most relevant for what follows in this paper, the input to ReLERNN is comprised of the frequency of each SNP, either as the minor or derived allele frequency. A collection of allele frequencies at multiple sites can be used to construct an allele frequency spectrum, which is simply a summary of the various frequencies within a sample. Adrión et al. (2020) showed that their accuracy in estimating  $\rho$  increased with more accurate estimates of allele frequencies, suggesting that these data are a key input.

Here, we examine the “black box” at the heart of ML estimates of recombination from pooled sequencing data. We propose that the allele frequency spectrum can be related to the population recombination parameter,  $\rho$ , and we provide a genealogical explanation for this relationship. We first demonstrate this connection using simulations. We then develop a simple ML model for inferring  $\rho$  from pooled sequencing data, showing that it is both accurate and robust to many assumptions. We apply our model, called NoDEAR (No Disequilibrium Estimation of Accurate Recombination), to data from humans, demonstrating that it is highly correlated with estimates using composite likelihood methods. Together, our investigations shed light into novel ways that recombination can affect the allele frequency spectrum, and how ML methods can help to uncover fundamental biological relationships.

## Genealogical effects on the allele frequency spectrum

The allele frequency spectrum is a central concept in modern population genetics. For a sample of  $n$  haploid chromosomes, we define the allele frequency spectrum as a vector of length  $n-1$  when considering derived allele frequencies and of length  $n/2$  when considering minor allele frequencies (rounding down if  $n$  is an odd number). For the derived frequency spectrum, the entries in the vector correspond to either the count or the proportion of all polymorphisms in a dataset found on 1, 2, 3, ...,  $n-1$  chromosomes. Elements of each vector therefore represent the fraction of all variants found at sample frequencies  $1/n$ ,  $2/n$ ,  $3/n$ , ...,  $(n-1)/n$ . Such an object is perhaps not an obvious source of information about recombination. One reason for this is that the expected frequency spectrum has been derived using multiple approaches (Ewens 1979; Tajima 1989; Fu 1995; Griffiths and Tavaré 1998;



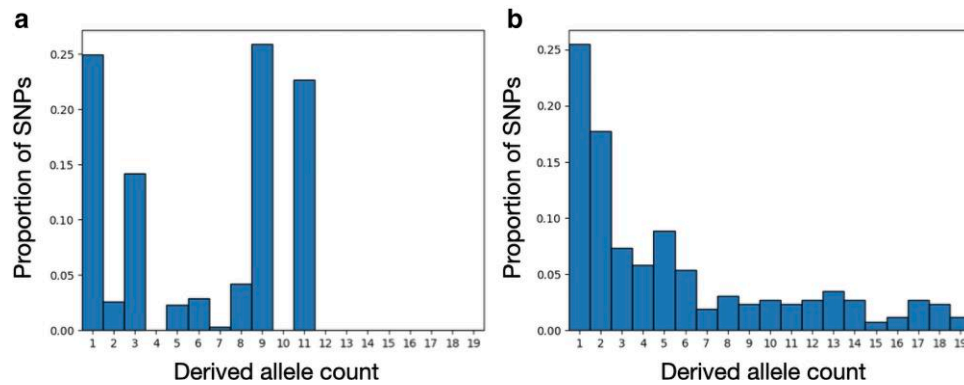
**Fig. 1.** Example gene trees, mutations, and allele frequency spectra. a) The top shows a hypothetical gene tree with  $n = 4$  tips and  $S = 3$  mutations (circles). Two mutations are of size 1 (have one descendant each) and one mutation is of size 2. The bottom shows the allele frequency spectrum that would result from this tree and mutations. b) The top again shows a gene tree and mutations. There are 2 mutations of size 1 and one mutation of size 3. Importantly, under the infinite sites model, mutations of size 3 are only possible on this gene tree topology, not on the one in a (mutations of size 2 are also possible in this tree, but none is shown). The bottom shows the resulting allele frequency spectrum.

Hudson 2015) and is the same across population sizes and recombination rates. However, these results are all expectations in the limit of an infinite number of loci; as we explain next, it is exactly the violation of this assumption that provides a link to recombination.

To understand how the allele frequency spectrum can tell us about recombination, consider the spectrum arising from a single non-recombining locus (Fig. 1a). At a single locus, there is a single tree topology, which limits the possible allele frequencies observed. Assuming an infinite sites model (i.e. the same mutation does not appear more than once), mutations can have only a limited set of frequencies: those possible in the local tree topology. Consider the toy example in Fig. 1a: only polymorphisms of “size” 1 and 2 are possible, since mutations can occur only on branches with either 1 or 2 descendants (Fu 1995). Given this, only polymorphisms with a frequency of 25% or 50% are possible. In contrast, in the tree shown in Fig. 1b, mutations of size 1, 2, or 3 are possible. Therefore, a different allele frequency spectrum can be produced by this tree topology.

In general, any particular topology sampled at a non-recombining locus will permit only a subset of allele frequencies, and this subset is likely to differ among trees at different loci. Even if mutations of the same size are permitted on 2 trees—for instance, if they have the same hierarchical topology—differences in branch lengths can still result in different overall allele frequency spectra due to different numbers of mutations of each size (Ferretti et al. 2013). Importantly, we have proved in the Appendix (<https://doi.org/10.5281/zenodo.14775487>) that no single topology can possibly give the allele frequency spectrum expected in the infinite-locus limit when  $n \geq 4$ .

We propose that it is exactly the sparsity of the allele frequency spectrum from a limited number of tree topologies that provides



**Fig. 2.** Simulated allele frequency spectra under different levels of recombination. a) The allele frequency spectrum produced for a sample of size  $n = 20$  in a simulated 50-kb region with low recombination ( $\rho = 0.282$ ). There were 2 different tree topologies found in this region. b) The allele frequency spectrum produced for a sample of size  $n = 20$  in a simulated 50-kb region with high recombination ( $\rho = 7711.7$ ). There were 3515 different tree topologies found in this region.

information about recombination. Recombining regions containing a small number of tree topologies will give sparse spectra, while regions with a large number of topologies will give smoother spectra because they are the sum across the spectra produced by each marginal topology. Most recombination events in the history of a sample will result in an additional tree, such that  $R$  recombination events in a genomic region can lead to at most  $R + 1$  topologies in the region (Hudson 1983; Griffiths and Marjoram 1996; Wiuf and Hein 1999; McVean and Cardin 2005; Marjoram and Wall 2006). Although not all recombination events will change the topology or branch lengths of a tree (Marjoram and Wall 2006; Ferretti et al. 2013), it is clear that the number of unique trees in a region will also be associated with  $\rho$  (Hudson and Kaplan 1985). If the number of marginal trees in a region is related to the allele frequency spectrum in any sort of straightforward manner, we should be able to use this representation of the data to estimate  $\rho$ . This interpretation is also consistent with the observation by Adrion et al. (2020) that the accuracy of ReLERNN increases with higher accuracy of allele frequency estimates, as the latter will of course also make the estimate of the allele frequency spectrum more accurate.

## Results

### Simulations connect recombination to the allele frequency spectrum

To test whether a relationship between recombination and the allele frequency spectrum exists, we carried out simulations in msprime (Kelleher et al. 2016). Comprehensive simulations across parameter space were carried out as described below; here, our goal was to simply demonstrate this link. We simulated  $n = 20$  haploid samples from the equivalent of a 50 kb region with constant population size,  $N = 70,000$ , and mutation rate per generation,  $\mu = 1.0 \times 10^{-8}$  (assuming an infinite sites mutation model).

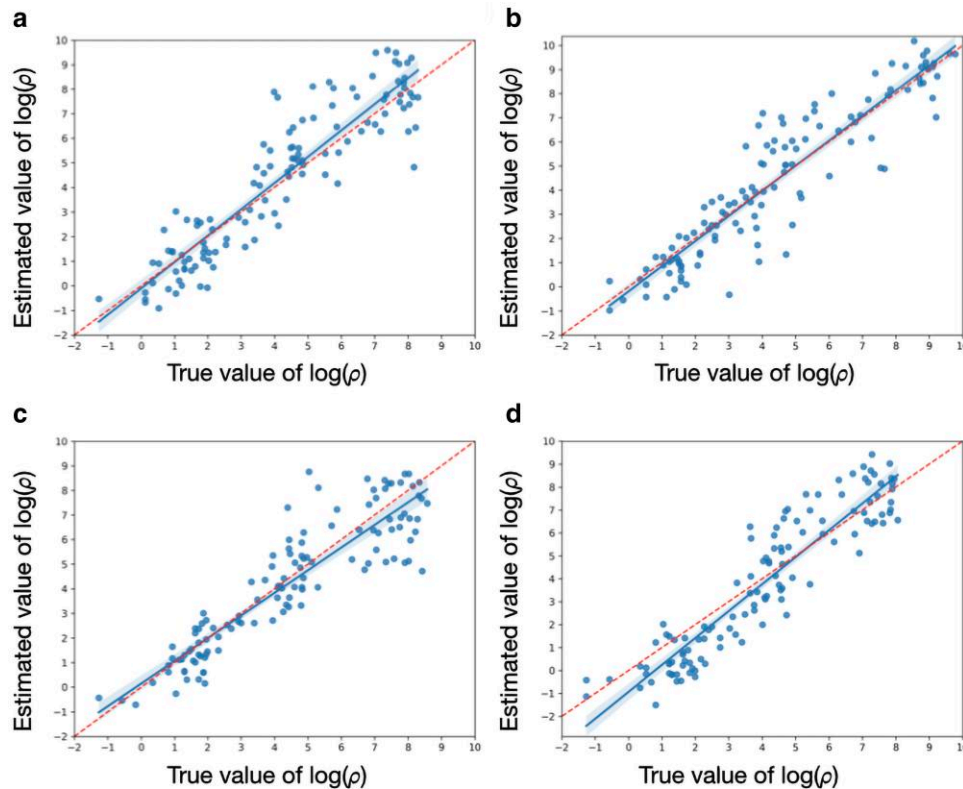
Figure 2 shows examples of 2 typical allele frequency spectra from simulations with low recombination ( $\rho = 0.282$ ) and high recombination ( $\rho = 7711.7$ ). As can be seen, the spectrum from a region with less recombination in its history (Fig. 2a) is multimodal, choppy, and hardly resembles the neutral expectation, while the spectrum from a region with high recombination (Fig. 2b) is smoother, approaching that of the expected spectrum. For reference, we also recorded the number of unique tree topologies in the 2 simulations: there are 2 trees in the low recombination condition and 3,515 trees in the high recombination condition.

The qualitative descriptions of spectra as “choppy” or “smooth” are not easy to summarize across thousands of simulations and would require that we assess each allele frequency spectrum visually. Instead, we attempt to capture these patterns quantitatively by calculating the Euclidean distance ( $L^2$ -norm) between the observed and expected allele frequency spectrum in each simulated window (as calculated using NumPy; Harris et al. 2020). Using this approach, genomic windows with smooth spectra will have  $L^2$ -norm = 0 and with highly irregular spectra will have  $L^2$ -norm  $\gg 0$ .  $L^2$ -norm can be calculated quickly for any dataset for use in quantitative comparisons. Although we use the Euclidean distance from the neutral-equilibrium expectation here, we describe how equivalent calculations can be carried out for non-equilibrium histories in the Discussion. Other standard measures of distance could be used (e.g. Kullback-Leibler divergence), though these may not be as flexible in non-equilibrium scenarios. For the examples used above, the values of  $L^2$ -norm are 0.347 in low recombination (Fig. 2a) and 0.069 in high recombination (Fig. 2b).

Given this framework, we simulated 10,000 recombining regions with the same parameters as above, using values of  $\rho$  ranging across 4 orders of magnitude. We found a highly negative correlation between  $\rho$  and  $L^2$ -norm calculated from the derived frequency spectrum (Spearman’s  $r_s = -0.90$ ; Pearson’s  $r_{xy} = -0.89$ ), such that higher recombination resulted in smoother spectra and lower Euclidean distances (Supplementary Fig. 1a). If we use the minor allele frequency spectrum instead, the correlation between  $\rho$  and  $L^2$ -norm is slightly reduced (Spearman’s  $r_s = -0.84$ ; Pearson’s  $r_{xy} = -0.82$ ; Supplementary Fig. 1b); this weaker relationship is expected given that there is less information in the minor spectrum. Regardless of which we use, as predicted, the allele frequency spectrum contains information about the amount of recombination in a sample. We therefore next develop a machine learning model that can predict  $\rho$  from this spectrum.

### NoDEAR: a machine learning model

In light of the above relationships, we developed a simplified ML model to estimate the recombination rate using only the allele frequency spectrum. Our goal is not to compete with ReLERNN, but instead to see how little data we can use and still accurately estimate  $\rho$ . Here, we explain how we trained our model, which we call NoDEAR (No Disequilibrium Estimation of Accurate Recombination), and following this we test its accuracy and robustness in multiple ways. The goal of NoDEAR is to highlight the power of an ML approach to recombination estimation that



**Fig. 3.** Correlation between true values of  $\rho$  and values estimated by NoDEAR. a) Correlation between values simulated under an equilibrium population history with no gene conversion (Spearman's  $r_s = 0.922$  and  $R^2 = 0.806$ ). Each dot represents one simulated 50-kb region. Blue lines represent best-fit regressions (as calculated in Seaborn; Waskom 2021), plus confidence intervals. Red dashed lines show the  $y = x$  line. b) Correlation between values simulated under non-equilibrium population histories with no gene conversion (Spearman's  $r_s = 0.925$ ;  $R^2 = 0.789$ ). c) Correlation between values simulated under a history with population structure and no gene conversion (Spearman's  $r_s = 0.887$ ;  $R^2 = 0.760$ ). d) Correlation between values simulated under an equilibrium population history with gene conversion (Spearman's  $r_s = 0.919$ ;  $R^2 = 0.703$ ).

uses solely the allele frequency spectrum; researchers interested in estimating  $\rho$  from empirical datasets should use alternative software.

NoDEAR uses XGBoost (Chen and Guestrin 2016) as implemented in Python to learn relationships between the allele frequency spectrum and  $\rho$ . The standard input to NoDEAR is the allele frequency spectrum represented as a vector of normalized proportions, such that the sum of all entries equals 1. By normalizing the values, we remove information about the number of SNPs in a dataset. This representation also obviously contains no information about the relative genomic positions of SNPs. The output of NoDEAR is a predicted value of  $\rho$ .

Training our ML model was straightforward. The reference simulations were again carried out using msprime (Kelleher et al. 2016) as described in the previous section. Later, we change the population history and size of genomic windows to test the effects of each. For training, we ran simulations with  $c$  varying across 5 orders of magnitude (from  $0.5 \times 10^{-12}$  to  $0.5 \times 10^{-7}$ ). We divided this range of recombination rates into 10 equally sized bins with 1000 simulated datasets in each, for a total of 10 000 simulated datasets. For each simulated recombining region, we recorded the number of recombination events,  $R$ , then calculated  $\rho$  by using the relationship  $\rho = R/a$ , where  $a$  is the harmonic series from 1 to  $n-1$  (Hudson and Kaplan 1985). While this value of  $\rho$  uses the expected tree length at a locus—an expectation that may not hold for shorter regions or in non-equilibrium populations—the information contained in  $R$  appears to be much more important for prediction. The value of  $\rho$  and the allele frequency vector for each region were passed to XGBoost for training.

Training on all 10,000 simulated datasets took 41.36 seconds on 1 core of an Intel Xeon processor with 100 Gb of available RAM, run on the Indiana University Research Desktop. We assessed the accuracy of the NoDEAR model on the training data by carrying out 5-fold cross-validation (using Scikit-learn; Pedregosa et al. 2011), with an average score of 0.87.

### Accuracy of recombination estimation from the allele frequency spectrum

After training NoDEAR on allele frequency spectra associated with a wide range of recombination histories, we asked how well this approach could predict  $\rho$  in new simulations not used in training. Reference simulations on 120 regions were carried out as described above, but with  $c$  varying across only 3 orders of magnitude (from  $10 \times 10^{-11}$  to  $10 \times 10^{-8}$ ), and then passed to NoDEAR for prediction. Runtime for prediction on the entire test dataset was 0.53 seconds. The correlation between the true value of  $\rho$  and the value estimated by NoDEAR was quite high, with Spearman's  $r_s = 0.922$  (Fig. 3a; calculated using SciPy; Virtanen et al. 2020). The predicted values of  $\rho$  also explain a large amount of the variation in the true values of  $\rho$  ( $R^2 = 0.806$ ).

To provide some context as to how accurate NoDEAR is compared to the state-of-the-art estimates of population recombination parameters, we applied the composite likelihood method implemented in the program, pyrho (Spence and Song 2019). We provided pyrho with the phased haplotypes from all 20 simulated individuals at each 50-kb region, as it cannot use just the allele frequency spectrum. Pyrho outputs the recombination rate per base per generation,  $c$ , rather than  $\rho$  ( $=4N_e c$ ), by assuming that the



demographic model inferred by this method is the only cause of genealogical heterogeneity (see Discussion). Regardless, the rank correlation between the inferred  $c$  from *pyrho* and the true  $\rho$  should be comparable to our results with NoDEAR.

Using *pyrho* to estimate parameters using the same test data, the correlation with the true values was extremely high, with Spearman's  $r_s = 0.97$  (Supplementary Fig. 2; we first removed estimates of  $c$  below  $10^{-20}$ , as there was a cluster of outlier points with  $c \approx 10^{-50}$ ). Although these results are more accurate than those using NoDEAR, they also took much longer to estimate (runtime = 2549.4 s). In addition, increasing the sample size used by NoDEAR to  $n = 100$  further increases the accuracy of inference (Spearman's  $r_s = 0.961$ ;  $R^2 = 0.893$ ), as the larger sample results in a more highly resolved allele frequency spectrum. This increase in sample size does not increase either the time to train or to test NoDEAR relative to  $n = 20$  results.

## Robustness of recombination estimation from the allele frequency spectrum

The above tests were all carried out under relatively simple conditions. We used a machine learning model trained and tested on an equilibrium population history, using the derived allele frequency spectrum, at only 1 window size (50 kb), and with no gene conversion (which could bias estimates of  $\rho$ ; Setter et al. 2022; Dutheil 2024). We therefore wanted to understand how robust our inferences were to additional model complexities.

We first trained and tested NoDEAR on the minor allele frequency spectrum, keeping all other conditions the same. As expected, this approach did worse than when using the derived spectrum, but only slightly so (Spearman's  $r_s = 0.867$ ;  $R^2 = 0.676$ ; Supplementary Table 1). We also carried out training and testing using additional genomic window sizes, including 10 kb, 100, 150, and 200 kb (keeping the per-base recombination rate the same in each). Across these window sizes, we found a range of correlations between the estimated and true values of  $\rho$  (Spearman's  $r_s$ : 0.823, 0.914, 0.906, and 0.931, respectively; Supplementary Table 1). Although the smallest window size (10 kb) shows some reduction in predictive accuracy (e.g. higher MSE; Supplementary Table 1)—possibly because there are not enough recombination events to distinguish among different values of  $\rho$ , or not enough SNPs to construct an accurate spectrum—NoDEAR behaves well at larger window sizes. However, we expect that at some larger window size predictive accuracy must go down, as every window will have an indistinguishably large number of recombination events and therefore indistinguishable values of  $\rho$ . These results also do not distinguish between the amount of recombination in a region and the number of SNPs in a region, the latter of which will also affect the accuracy of the allele frequency spectrum. The exact physical scale over which NoDEAR will be effective will therefore be a function of  $c$  and  $\mu$  (the per-generation mutation rate), and will differ among biological systems.

We assessed the effect of non-equilibrium demographic conditions by generating test data with such histories, but using a NoDEAR model that was trained on equilibrium conditions. We simulated 2 types of non-equilibrium demographies: in the first set, we sampled changes in population size through time by first drawing a number of times at which a population changed in size from a Poisson distribution with rate parameter,  $\lambda = 3$ . This means that most simulated loci will have changed in size 3 times, but some will have changed 2 or 4 (or 1 or 5, etc.) times as well. For each change in size, we then drew a new size from a normal distribution centered on  $N = 70,000$ . These simulations therefore capture expansions and contractions in a single population over

time, independently carried out for each recombining region. In the second set of non-equilibrium histories, we simulated population structure by having 2 sub-populations that split 10,000 generations ago for all loci (all populations have  $N = 70,000$ ). The allele frequency spectrum was constructed by sampling  $n = 10$  individuals from each of the 2 sub-populations and combining them for a total of  $n = 20$ . In both sets of non-equilibrium simulations, the true value of  $\rho$  was again calculated via the number of recombination events in the history of a region by using the relationship  $\rho = R/a$ .

Perhaps surprisingly (see Discussion), we observed no reduction in accuracy of NoDEAR when predicting  $\rho$  in populations whose sizes are changing over time, even when our model is trained on equilibrium histories. For 50-kb windows, results from test datasets of non-equilibrium histories are essentially the same as with equilibrium histories (Spearman's  $r_s = 0.925$ ;  $R^2 = 0.789$ ; Fig. 3b). Supplementary Table 1 shows that while inference on non-equilibrium populations is not better across all window sizes, neither does it get much worse than results from equilibrium populations. Likewise, simulations with a history of population subdivision led to slightly worse results in 50-kb regions (Spearman's  $r_s = 0.887$ ;  $R^2 = 0.760$ ; Fig. 3c), but no consistent reduction across other window sizes (Supplementary Table 1).

Finally, we generated a test dataset in which both crossing-over and gene conversion can occur (all previous simulations only had crossing-over). Many LD-based methods for estimating  $\rho$  cannot capture the effects of both types of recombination, sometimes overestimating and sometimes underestimating (e.g. *pyrho*) the total recombination rate (Dutheil 2024). In these simulations, gene conversion events make up 50% of all recombination events, with tract length 300 bp. Again, however, we see no reduction in the accuracy of NoDEAR, even though it was trained on data with no gene conversion (50-kb: Spearman's  $r_s = 0.919$ ;  $R^2 = 0.703$ ; Fig. 3d; Supplementary Table 1 contains all other window sizes). NoDEAR is accurately predicting the amount of recombination, regardless of the exact mechanism of recombination.

## Application to data from humans

To further demonstrate the use of the allele frequency spectrum as a method for estimating the population recombination parameter, we applied NoDEAR to a dataset from humans. Because we do not know the true recombination rates across loci in these data, we also ran *pyrho* for comparison. We chose 10 diploid samples from Finland ( $n = 20$ ), using only SNPs from chromosome 6, giving a total of 3,408 50-kb windows (The 1000 Genomes Project Consortium 2015).

One issue that occurs in real data that did not occur in our simulations is missing genotypes. To deal with missing data—which can result in different counts of the minor or derived allele at different sites—we used a vector with counts of SNPs in 10 bins of allele frequencies (e.g. 0–0.05, 0.05–0.10, etc.). Because alleles were not assigned as ancestral or derived in the human data, we used NoDEAR trained on the minor allele frequency spectrum. We only used 50-kb windows of the genome that contained at least 100 SNPs, to ensure that there was enough data for estimation. In total, this resulted in 2,602 50-kb windows that could be analyzed by NoDEAR.

NoDEAR ran quickly on all 2,602 loci, taking 25.4 s; in comparison *pyrho* took 55,100.2 s (i.e. 15.3 h), including steps that inferred the population demography of the Finnish sample (which only took 156.4 s). The value of  $\rho$  inferred by NoDEAR varied from 0.172 to 2453.12 across chromosome 6; *pyrho* reports  $c$  for the same data, ranging from  $4.64 \times 10^{-11}$  to  $1.51 \times 10^{-7}$  (taking the

average of each window; we again removed 104 values with  $c \approx 10^{-50}$ ). Overall, there were 2,509 50-kb windows with estimates from both NoDEAR and pyrro. The correlation between the 2 estimates of recombination among these windows was quite high (Spearman's  $r_s = 0.60$ ; [Supplementary Fig. 3](#)), though maybe not as high as anticipated.

One biological feature that NoDEAR does not consider are recombination hotspots, which are common in humans ([McVean et al. 2004](#)). Hotspots result in intense crossing-over in a short region, with little crossing-over nearby; such recombination may not leave as strong a signal on the allele frequency spectrum. To test this idea indirectly, we took the 261 genomic windows in the human data with the highest internal variance in recombination rate (as predicted by pyrro), assuming that this high variance was a sign of possible hotspots. The correlation between the estimates of recombination from NoDEAR and pyrro for these windows was indeed lower (Spearman's  $r_s = 0.44$ ), possibly indicating that NoDEAR is not doing as well at predicting recombination when it is highly punctate. Nevertheless, our results overall provide further evidence that one can infer the recombination history of natural populations using only the allele frequency spectrum.

## Discussion

Recombination is of interest to a wide variety of biologists but has an especially important role in evolutionary biology because of its role in moderating the influence of natural selection (e.g. [Hill and Robertson 1966](#)). As a result, there are many approaches for estimating recombination across the genome. One very common approach uses polymorphism data to estimate the amount of recombination in the history of a small sample, resulting in an estimate of the population recombination parameter,  $\rho$  (sometimes called  $C$ ). This estimate represents an integrated history of recombination over time and across individuals that have left descendants in a sample. Because most modern methods that estimate  $\rho$  use LD among sites, often population-based estimates of recombination are simply called “LD-based” estimates. It should be noted, however, that some of the first statistical estimators of  $\rho$  used the variance in pairwise differences between phased haplotypes, not LD ([Hudson 1987](#); [Wakeley 1997](#)).

Regardless of the exact approach used, most previous model-based methods for estimating  $\rho$  required that the genotypes of the individuals being considered could be associated with those individuals. Sometimes individual alleles are arranged along chromosomes within individuals (i.e. gametic LD), and sometimes diploid genotypes along chromosomes are associated with individuals (i.e. genotypic or zygotic LD). Here, we have demonstrated that the allele frequency spectrum alone can be used to estimate population recombination rates. The allele frequency spectrum is a vector of SNP counts or proportions at each frequency in a sample and contains no information about alleles or genotypes at different loci found in any particular individual. However, we find that the allele frequency spectrum does indirectly contain information about the number of marginal gene trees in a region ([Figs. 1 and 2](#)). Because the number of gene trees reflects the number of recombination events, the spectrum can be used in a straightforward way to estimate  $\rho$ . Similar ideas were also used explicitly by [Beeravolu et al. \(2018\)](#) and implicitly by [Burger et al. \(2022\)](#) to estimate recombination.

Our exploration of the role of the allele frequency spectrum in recombination estimation was inspired by the software ReLERNN ([Adron et al. 2020](#)). ReLERNN is a machine learning method that

can infer  $\rho$  from either genotype data or pooled sequencing data, carrying out both tasks with high accuracy. As pooled sequencing data does not contain any information on genotypes of individuals, we were curious about the “black box” at the heart of the ReLERNN machine learning model. We reasoned that the model was likely using the allele frequency spectrum to predict  $\rho$  and indeed our analyses suggest that this is the case. Importantly, however, ReLERNN could be using additional information not considered by the simplified model learned by our software, NoDEAR. We purposefully removed information about both the number of SNPs in a window and the location of SNPs in each window. In the Introduction, we discussed how the number of SNPs could be informative about  $\rho$  (at least in non-neutral scenarios), but the location of SNPs may be even more informative, including in neutral scenarios. One could imagine calculating the correlation of the allele frequency spectrum (or other measures of variation) among sub-windows of a larger region in order to see how quickly it changes. Such information is surely informative about  $\rho$  and may be being used by ReLERNN. Indeed, ReLERNN did slightly better than NoDEAR at predicting recombination rates on our reference simulations (Spearman's  $r_s = 0.93$ ). As our goal was largely to understand the information contained within the allele frequency spectrum—and neither to fully dissect ReLERNN nor to build a competitor software to it—we did not explore the possibilities contained within these other pieces of data further.

In addition to demonstrating that the allele frequency spectrum can be used to estimate recombination, we also showed that a very reduced representation of this spectrum could be used for the same purpose. We calculated the Euclidean distance ( $L^2$ -norm) between the allele frequency spectrum from a region and the expected equilibrium spectrum, showing that this was highly predictive of  $\rho$  ([Supplementary Fig. 1](#)). This simple summary statistic works well because it captures the main effect of recombination in a region: that the spectrum generated by summing over many trees will be much smoother and therefore closer in distance to the expected spectrum, than the spectrum from a small number of trees. We could of course have trained a machine learning model using this distance, but as it is a single value, we would not do much better than the rank correlation between  $L^2$ -norm and  $\rho$ , which was already quite high (Spearman's  $r_s = -0.89$ ). How could one use  $L^2$ -norm if a population did not have an equilibrium history? One straightforward solution is to build a reference allele frequency spectrum by summing over data from all windows together. This spectrum has the maximal amount of recombination possible and therefore the Euclidean distance between individual windows and this reference should again be proportional to the recombination rate in each window.

One possibly surprising result from our simulations is that the inferences made by NoDEAR were not affected by non-equilibrium population histories—either changes in population size or structure—even when the model was trained on an equilibrium history. While non-equilibrium histories will have allele frequency spectra that have a different shape from the equilibrium spectrum produced during training, our results suggest that NoDEAR is not learning this shape per se, but rather the choppy/smoothness of spectra generated by different levels of recombination. This behavior is quite helpful, as it means that one would not have to retrain the model on every new dataset.

Our results concerning the lack of an effect of non-equilibrium histories on the estimation of recombination stands in apparent contrast to some previous results and require clarification. Multiple previous simulation studies have shown that non-equilibrium histories can affect estimates of the recombination rate ([McVean et al. 2002](#);

Smith and Fearnhead 2005; Johnston and Cutler 2012; Dapper and Payseur 2018; Spence and Song 2019; Adrion et al. 2020; Samuk and Noor 2022; Raynaud et al. 2023; Dutheil 2024). However, different studies often mean something slightly different by “recombination rate.” As mentioned earlier,  $\rho$  is the population recombination parameter, defined as  $4N_e c$ , such that both the population history in a region and the per-generation recombination rate in a region will affect the number of recombination events found in a sample. In an equilibrium population, estimates of  $N_e$  can be used to estimate  $c$  (sometimes called  $r$ ) directly. Even in populations with non-equilibrium histories, accurate estimates of  $N_e$ —taking into account this history—can be used to estimate  $c$  (Spence and Song 2019; Adrion et al. 2020). It is the estimate of  $c$  that can be biased when non-equilibrium histories are not taken into account (i.e. when the wrong value of  $N_e$  is used), not the amount of recombination that has occurred in a sample from nature (i.e.  $\rho$ ).

Estimating  $c$  from  $\rho$  assumes that the inferred demographic model is the only cause of genealogical heterogeneity. If natural selection acts to either reduce (e.g. positive selection or background selection) or increase (e.g. balancing selection) the height of a genealogy, the number of recombination events may no longer be proportional to  $c$  (Smith and Fearnhead 2005; O'Reilly et al. 2008; Spence and Song 2019; Adrion et al. 2020). This occurs because the effective population size at such loci is not determined solely by demographic history. We have chosen to construct NoDEAR to only predict  $\rho$  from data; it is likewise trained only on  $\rho$ -values directly calculated from the actual number of recombination events produced by our simulations. It is this relatively assumption-free approach that allows NoDEAR to be accurate under different non-equilibrium conditions (and mechanisms of recombination, such as gene conversion) for which it was not trained. We imagine it would do a similarly accurate job of estimating  $\rho$  in the presence of non-neutral evolution.

The allele frequency spectrum is a fundamental measurement of variability in DNA sequences, used for the inference of both selection and demography. Here, we have shown that it also contains evidence of recombination, as it encodes information about the number of marginal gene trees in a genomic window. Although this property of the allele frequency spectrum has rarely been recognized previously (e.g. Beeravolu et al. 2018), certainly there are many theoretical studies that are relevant to the utility of the spectrum for this purpose. For instance, work on the average distance between recombination events (e.g. Deng et al. 2021) and the average effect of recombination events on tree topologies (e.g. Ferretti et al. 2013) both provide important context for the power of the allele frequency spectrum alone to infer recombination. We hope that future work can further explore the application of these, or related, approaches.

## Data availability

All code used in this paper is available on GitHub (<https://github.com/smishra677/NoDEAR>).

Supplemental material available at GENETICS online.

## Acknowledgments

We thank Jeff Adrion and Andy Kern for discussion, the Kern-Ralph co-lab for helpful feedback at an early stage of this work, Daniel Rickert for co-authoring the appendix, and Richard Wang for comments on the manuscript. Konrad Lohse, Thomas Decroly, and one anonymous reviewer also provided constructive criticism.

## Funding

This work was supported by U.S. National Science Foundation grant DBI-2146866.

## Conflicts of interest

The authors declare no conflicts of interest.

## Literature cited

- Adrion JR, Galloway JG, Kern AD. 2020. Predicting the landscape of recombination using deep learning. *Mol Biol Evol.* 37(6):1790–1808. <https://doi.org/10.1093/molbev/msaa038>.
- Barroso GV, Puzović N, Dutheil JY. 2019. Inference of recombination maps from a single pair of genomes and its application to ancient samples. *PLoS Genet.* 15(11):e1008449. <https://doi.org/10.1371/journal.pgen.1008449>.
- Beeravolu CR, Hickerson MJ, Frantz LAF, Lohse K. 2018. ABLE: block-wise site frequency spectra for inferring complex population histories and recombination. *Genome Biol.* 19(1):145. <https://doi.org/10.1186/s13059-018-1517-y>.
- Bernett J, Blumenthal DB, Grimm DG, Haselbeck F, Joeres R, Kalinina OV, List M. 2024. Guiding questions to avoid data leakage in biological machine learning applications. *Nat Methods.* 21(8):1444–1453. <https://doi.org/10.1038/s41592-024-02362-y>.
- Burger KE, Pfaffelhuber P, Baumdicker F. 2022. Neural networks for self-adjusting mutation rate estimation when the recombination rate is unknown. *PLoS Comput Biol.* 18(8):e1010407. <https://doi.org/10.1371/journal.pcbi.1010407>.
- Chan AH, Jenkins PA, Song YS. 2012. Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genet.* 8(12):e1003090. <https://doi.org/10.1371/journal.pgen.1003090>.
- Chen T, Guestrin C. 2016. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; San Francisco, CA, United States. Association for Computing Machinery. p. 785–794.
- Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet.* 14(4):262–274. <https://doi.org/10.1038/nrg3425>.
- Dapper AL, Payseur BA. 2018. Effects of demographic history on the detection of recombination hotspots from linkage disequilibrium. *Mol Biol Evol.* 35(2):335–353. <https://doi.org/10.1093/molbev/msx272>.
- Deng Y, Song YS, Nielsen R. 2021. The distribution of waiting distances in ancestral recombination graphs. *Theor Popul Biol.* 141:34–43. <https://doi.org/10.1016/j.tpb.2021.06.003>.
- Dutheil JY. 2024. On the estimation of genome-average recombination rates. *Genetics.* 227(2):iyae051. <https://doi.org/10.1093/genetics/iyae051>.
- Ewens WJ. 1979. Testing the generalized neutrality hypothesis. *Theor Popul Biol.* 15(2):205–216. [https://doi.org/10.1016/0040-5809\(79\)90035-2](https://doi.org/10.1016/0040-5809(79)90035-2).
- Feder AF, Petrov DA, Bergland AO. 2012. LDx: estimation of linkage disequilibrium from high-throughput pooled resequencing data. *PLoS One.* 7(11):e48588. <https://doi.org/10.1371/journal.pone.0048588>.
- Ferretti L, Disanto F, Wiehe T. 2013. The effect of single recombination events on coalescent tree height and shape. *PLoS One.* 8(4):e60123. <https://doi.org/10.1371/journal.pone.0060123>.
- Flagel L, Brandvain Y, Schrider DR. 2019. The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Mol Biol Evol.* 36(2):220–238. <https://doi.org/10.1093/molbev/msy224>.

- Fu Y-X. 1995. Statistical properties of segregating sites. *Theor Popul Biol.* 48(2):172–197. <https://doi.org/10.1006/tpbi.1995.1025>.
- Gao F, Ming C, Hu W, Li H. 2016. New software for the fast estimation of population recombination rates (FastEPRR) in the genomic era. *G3 (Bethesda)*. 6(6):1563–1571. <https://doi.org/10.1534/g3.116.028233>.
- Griffiths RC, Marjoram P. 1996. Ancestral inference from samples of DNA sequences with recombination. *J Comput Biol.* 3(4):479–502. <https://doi.org/10.1089/cmb.1996.3.479>.
- Griffiths RC, Tavaré S. 1998. The age of a mutation in a general coalescent tree. *Stoch Model.* 14(1–2):273–295. <https://doi.org/10.1080/15326349808807471>.
- Hahn MW. 2018. *Molecular Population Genetics*. Sunderland (MA): Sinauer Associates.
- Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. 2020. Array programming with NumPy. *Nature*. 585(7825):357–362. <https://doi.org/10.1038/s41586-020-2649-2>.
- Haubold B, Pfaffelhuber P, Lynch M. 2010. mlRho—a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Mol Ecol.* 19(s1):277–284. <https://doi.org/10.1111/j.1365-294X.2009.04482.x>.
- Hermann P, Heissl A, Tiemann-Boege I, Futschik A. 2019. *LDjump*: estimating variable recombination rates from population genetic data. *Mol Ecol Resour.* 19(3):623–638. <https://doi.org/10.1111/1755-0998.12994>.
- Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res.* 8(3):269–294. <https://doi.org/10.1017/S0016672300010156>.
- Huang X, Rymbekova A, Dolgova O, Lao O, Kuhlwillm M. 2024. Harnessing deep learning for population genetic inference. *Nat Rev Genet.* 25(1):61–78. <https://doi.org/10.1038/s41576-023-00636-3>.
- Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol.* 23(2):183–201. [https://doi.org/10.1016/0040-5809\(83\)90013-8](https://doi.org/10.1016/0040-5809(83)90013-8).
- Hudson RR. 1987. Estimating the recombination parameter of a finite population model without selection. *Genet Res.* 50(3):245–250. <https://doi.org/10.1017/S0016672300023776>.
- Hudson RR. 2001. Two-locus sampling distributions and their application. *Genetics*. 159(4):1805–1817. <https://doi.org/10.1093/genetics/159.4.1805>.
- Hudson RR. 2015. A new proof of the expected frequency spectrum under the standard neutral model. *PLoS One.* 10(7):e0118087. <https://doi.org/10.1371/journal.pone.0118087>.
- Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*. 111(1):147–164. <https://doi.org/10.1093/genetics/111.1.147>.
- Johnston SE. 2024. Understanding the genetic basis of variation in meiotic recombination: past, present, and future. *Mol Biol Evol.* 41(7):msae112. <https://doi.org/10.1093/molbev/msae112>.
- Johnston HR, Cutler DJ. 2012. Population demographic history can cause the appearance of recombination hotspots. *Am J Hum Genet.* 90(5):774–783. <https://doi.org/10.1016/j.ajhg.2012.03.011>.
- Kamm JA, Spence JP, Chan J, Song YS. 2016. Two-locus likelihoods under variable population size and fine-scale recombination rate estimation. *Genetics*. 203(3):1381–1399. <https://doi.org/10.1534/genetics.115.184820>.
- Kelleher J, Etheridge AM, McVean G. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol.* 12(5):e1004842. <https://doi.org/10.1371/journal.pcbi.1004842>.
- Korfmann K, Gaggiotti OE, Fumagalli M. 2023. Deep learning in population genetics. *Genome Biol Evol.* 15(2):evad008. <https://doi.org/10.1093/gbe/evad008>.
- Lin K, Futschik A, Li H. 2013. A fast estimate for the population recombination rate based on regression. *Genetics*. 194(2):473–484. <https://doi.org/10.1534/genetics.113.150201>.
- Marjoram P, Wall JD. 2006. Fast “coalescent” simulation. *BMC Genet.* 7(1):16. <https://doi.org/10.1186/1471-2156-7-16>.
- McVean GAT. 2002. A genealogical interpretation of linkage disequilibrium. *Genetics*. 162(2):987–991. <https://doi.org/10.1093/genetics/162.2.987>.
- McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*. 160(3):1231–1241. <https://doi.org/10.1093/genetics/160.3.1231>.
- McVean GAT, Cardin NJ. 2005. Approximating the coalescent with recombination. *Philos Trans R Soc B.* 360(1459):1387–1393. <https://doi.org/10.1098/rstb.2005.1673>.
- McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science*. 304(5670):581–584. <https://doi.org/10.1126/science.1092500>.
- O'Reilly PF, Birney E, Balding DJ. 2008. Confounding between recombination and selection, and the ped/pop method for detecting selection. *Genome Res.* 18(8):1304–1313. <https://doi.org/10.1101/gr.067181.107>.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. 2011. Scikit-learn: machine learning in python. *J Mach Learn Res.* 12:2825–2830.
- Peñalba JV, Wolf JBW. 2020. From molecules to populations: appreciating and estimating recombination rate variation. *Nat Rev Genet.* 21(8):476–492. <https://doi.org/10.1038/s41576-020-0240-1>.
- Raynaud M, Gagnaire P-A, Galtier N. 2023. Performance and limitations of linkage-disequilibrium-based methods for inferring the genomic landscape of recombination and detecting hotspots: a simulation study. *Peer Community J.* 3:e27. <https://doi.org/10.24072/pcjournal.254>.
- Samuk K, Noor MAF. 2022. Gene flow biases population genetics inference of recombination rate. *G3 (Bethesda)*. 12(11):jkac236. <https://doi.org/10.1093/g3journal/jkac236>.
- Schlötterer C, Tobler R, Kofler R, Nolte V. 2014. Sequencing pools of individuals—mining genome-wide polymorphism data without big funding. *Nat Rev Genet.* 15(11):749–763. <https://doi.org/10.1038/nrg3803>.
- Schrider DR, Kern AD. 2018. Supervised machine learning for population genetics: a new paradigm. *Trends Genet.* 34(4):301–312. <https://doi.org/10.1016/j.tig.2017.12.005>.
- Setter D, Ebdon S, Jackson B, Lohse K. 2022. Estimating the rates of crossover and gene conversion from individual genomes. *Genetics*. 222(1):iyac100. <https://doi.org/10.1093/genetics/iyac100>.
- Smith NGC, Fearnhead P. 2005. A comparison of three estimators of the population-scaled recombination rate: accuracy and robustness. *Genetics*. 171(4):2051–2062. <https://doi.org/10.1534/genetics.104.036293>.
- Spence JP, Song YS. 2019. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Sci Adv.* 5(10):eaaw9206. <https://doi.org/10.1126/sciadv.aaw9206>.
- Stevenson LS, Woerner AE, Kidd JM, Kelley JL, Veeramah KR, McManus KF, Project GAG, Bustamante CD, Hammer MF, Wall JD. 2016. The



- time scale of recombination rate evolution in great apes. *Mol Biol Evol.* 33(4):928–945. <https://doi.org/10.1093/molbev/msv331>.
- Sved JA. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor Popul Biol.* 2(2):125–141. [https://doi.org/10.1016/0040-5809\(71\)90011-6](https://doi.org/10.1016/0040-5809(71)90011-6).
- Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics.* 123(3):585–595. <https://doi.org/10.1093/genetics/123.3.585>.
- The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature.* 526(7571):68–74. <https://doi.org/10.1038/nature15393>.
- Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat Methods.* 17(3):261–272. <https://doi.org/10.1038/s41592-019-0686-2>.
- Wakeley J. 1997. Using the variance of pairwise differences to estimate the recombination rate. *Genet Res.* 69(1):45–48. <https://doi.org/10.1017/S0016672396002571>.
- Wall JD. 2000. A comparison of estimators of the population recombination rate. *Mol Biol Evol.* 17(1):156–163. <https://doi.org/10.1093/oxfordjournals.molbev.a026228>.
- Waskom M. 2021. Seaborn: statistical data visualization. *J Open Source Softw.* 6(60):3021. <https://doi.org/10.21105/joss.03021>.
- Weir BS. 1979. Inferences about linkage disequilibrium. *Biometrics.* 35(1):235–254. <https://doi.org/10.2307/2529947>.
- Weir BS, Hill WG. 1986. Nonuniform recombination within the human  $\beta$ -globin gene cluster. *Am J Hum Genet.* 38:776–781.
- Wu C, He J. 1999. Recombination as a point process along sequences. *Theor Popul Biol.* 55(3):248–259. <https://doi.org/10.1006/tpbi.1998.1403>.

Editor: K. Lohse