bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-4.0 International license.

Inferring the landscape of recombination using recurrent neural networks

Jeffrey R. Adrion[1,†], **Jared G. Galloway**[1,†], **Andrew D. Kern**[1]

[1]Institute of Ecology and Evolution, University of Oregon

***For correspondence:**
jadrion@uoregon.edu

[†]These authors contributed equally to this work

Abstract Accurately inferring the genome-wide landscape of recombination rates in natural populations is a central aim in genomics, as patterns of linkage influence everything from genetic mapping to understanding evolutionary history. Here we describe ReLERNN, a deep learning method for accurately estimating a genome-wide recombination landscape using as few as four samples. Rather than use summaries of linkage disequilibrium as its input, ReLERNN considers columns from a genotype alignment, which are then modeled as a sequence across the genome using a recurrent neural network. We demonstrate that ReLERNN improves accuracy and reduces bias relative to existing methods and maintains high accuracy in the face of demographic model misspecification. We apply ReLERNN to natural populations of African *Drosophila melanogaster* and show that genome-wide recombination landscapes, while largely correlated among populations, exhibit important population-specific differences. Lastly, we connect the inferred patterns of recombination with the frequencies of major inversions segregating in natural *Drosophila* populations.

Introduction

Recombination plays an essential role in the meiotic production of gametes in most sexual species, and is often required for proper pairing and segregation of chromosomes (***Hunter et al., 2006***; ***Mather, 1938***; ***Smith and Nicolas, 1998***). During meiotic recombination, double-strand breaks are resolved as crossover or non-crossover recombination events along the chromosome, and as such, homologous chromosomes can exchange genetic information (reviewed in ***Kirkpatrick, 2010***; ***Zelkowski et al., 2019***). Thus while recombination is often critical to development and reproduction, it also has profound effects on both evolutionary and population genomics (***Burt, 2000***; ***Felsenstein, 1974***; ***Haenel et al., 2018***; ***Hartfield and Otto, 2011***; ***Hill and Robertson, 1966***; ***Kondrashov, 1982***).

Indeed, the population recombination rate $\rho = 4Nr$ is a central parameter in population and statistical genetics (reviewed in ***Hahn, 2018***), as ρ largely determines patterns of linkage disequilibrium (LD) across the genome. In regions of the genome where ρ is relatively small we expect increased levels of LD, and conversely in genomic compartments with high ρ we expect little LD. Deviations from our expected levels of LD given the local recombination rate can be illustrative of the influence of other evolutionary forces such as selection or migration. For example, selective sweeps are expected to dramatically elevate LD near the target of selection (***Kim and Nielsen, 2004***; ***O'Reilly et al., 2008***; ***Parsch et al., 2001***).

Structural variation itself is expected to modulate the landscape of recombination along the chromosomes, as both crossovers and non-crossovers are predicated on the alignment of homologous sequences, and structural rearrangements may directly impact those alignments. Chromosomal inversions, long-known to suppress crossing over along a chromosome (e.g. ***Sturtevant, 1921***), are

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-4.0 International license.

perhaps the most well-studied example of such structural variation. Inversion polymorphisms have been implicated in diverse evolutionary phenomena including local adaptation (***Ayala et al., 2013***; ***Kirkpatrick and Barton, 2006***; ***Lowry and Willis, 2010***), reproductive isolation (***Ayala et al., 2013***; ***Noor et al., 2001***; ***Rieseberg, 2001***), and the maintenance of meiotic drive complexes (***Jaenike, 2001***; ***Presgraves et al., 2009***). As suppressors of recombination, we expect *a priori* that segregating inversions should show distinct histories of recombination in comparison to standard karyotype chromosomes.

While recombination plays a central role in meiosis and reproduction, the frequency and distribution of crossovers along the chromosomes are themselves phenotypes that can evolve (reviewed in ***Kirkpatrick, 2010***; ***Ritz et al., 2017***). Importantly, recombination rate variation exists between species, among sexes of the same species (males generally having shorter maps than females), and extends even between individuals of the same sex (***Kong et al., 2010***; ***Singh et al., 2013***; ***Winckler et al., 2005***). Yet while there is abundant variation in the rate of recombination within and between taxa, most methods for accurately measuring this variation involve painstaking experiments or large pedigrees. Thus genetics, as a field, would like to have a tool for directly estimating recombination rates from sequence data, without relying on pedigree genotyping or other ancillary information.

Accordingly, there is a rich history of estimating ρ in population genetics, including efforts to obtain minimum bounds on the number of recombination events (***Hudson and Kaplan, 1985***; ***Myers and Griffiths, 2003***; ***Wiuf, 2002***), methods of moments estimators (***Hudson, 1987***; ***Wakeley, 1997***), composite likelihood estimators (***Chan et al., 2012***; ***Hudson, 2002***; ***McVean et al., 2002***), and summary likelihood estimators (***Li and Stephens, 2003***; ***Wall, 2000***). Recently, supervised machine learning methods for estimating ρ have entered the fray (***Gao et al., 2016***; ***Lin et al., 2013***), and have proven to be competitive in accuracy with state-of-the-art composite likelihood methods such as LDhat (***McVean et al., 2002***), often with far less computing effort.

To this end, we sought to develop a novel method for inferring rates of recombination directly from a sequence alignment through the use of deep learning. In recent years deep artificial neural networks (ANNs) have produced remarkable performance gains in computer vision (***Krizhevsky et al., 2012***; ***Szegedy et al., 2015***), speech recognition (***Hinton et al., 2012***), natural language processing (***Sutskever et al., 2014***), and data preprocessing tasks such as denoising (***Vincent et al., 2008***). Perhaps most illustrative of the potential of deep learning is the remarkable success of convolutional neural networks (CNNs; ***Lecun et al., 1998***) on problems in image analysis. For example, prior to the introduction of CNNs to the annual ImageNet Large Scale Visual Recognition Challenge (***Krizhevsky et al., 2012***), no method had achieved an error rate of less than 25% on the ImageNet data set. In the years that followed, CNNs succeeded in reducing this error rate below 5%, exceeding human accuracy on the same tasks (***Russakovsky et al., 2015***).

In this study we focus our efforts on recurrent neural networks (RNNs), a promising network architecture for population genomics, which has proven adept for analyzing sequential data of arbitrary lengths (***Graves et al., 2013***). Unlike other machine learning methods, deep learning approaches do not require a predefined feature vector. When fed labeled training data (e.g. a set of haplotypes simulated under a known recombination rate), these methods algorithmically create their own set of informative statistics that prove most effective for solving the specified problem. By training deep learning networks directly on sequence alignments, we allow the neural network to automatically extract informative features from the data without human supervision. Learning directly from a sequence alignment for population genetic inference has recently been shown to be possible using CNNs (***Chan et al., 2018***; ***Flagel et al., 2018***), and as we show below, is also true for RNNs.

Here we introduce **Re**combination **L**andscape **E**stimation using **R**ecurrent **N**eural **N**etworks, an RNN-based method for estimating the genomic landscape of recombination rates directly from a phased genotype alignment. We found that ReLERNN is both highly accurate and out-performs competing methods at small sample sizes. We also show that ReLERNN retains its high accuracy in

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-4.0 International license.

the face of demographic model misspecification. We then apply ReLERNN to population genomic data from African samples of *Drosophila melanogaster*. We demonstrate that the landscape of recombination is largely conserved in this species, yet individual regions of the genome show marked population-specific differences. Finally, we found that chromosomal inversion frequencies directly impact the inferred rate of recombination, and we demonstrate that the role for inversions in suppressing recombination extends far beyond the inversion breakpoints themselves.

Results

ReLERNN: an accurate method for estimating the genome-wide recombination landscape

We developed ReLERNN, a new deep learning method for accurately predicting genome-wide per-base recombination rates from as few as four phased chromosomes. Briefly, ReLERNN provides an end-to-end inferential pipeline for estimating a recombination landscape from a population sample: it takes as input a user-filtered Variant Call Format (VCF) file of phased genotypes, and from this estimates a set of simulation parameters reflective of the input samples. ReLERNN then uses the coalescent simulation program, msprime (***Kelleher et al., 2016***), to simulate training, validation, and test data sets under either a user-supplied or an inferred demographic history, seeking to mimic population genetic properties of the empirical samples. ReLERNN trains a specific type of RNN, known as a Gated Recurrent Unit (GRU), to predict the per-base recombination rate for these simulations, using only the raw genotype matrix and a vector of genomic coordinates for each simulation example (***Figure 1***). It then uses this trained network to estimate genome-wide per-base recombination rates for empirical samples using a sliding-window approach. ReLERNN can optionally estimate 95% confidence intervals around each prediction using a parametric bootstrapping approach, and it uses these bootstrap estimates to correct for inherent biases in the training process (see Materials and Methods; ***Figure 1–Figure Supplement 1***).

A key feature of ReLERNN's network architecture is the bidirectional GRU layer (***Figure 1*** inlay), which takes advantage of the sequential nature of genomic data. While vanilla (feed-forward) networks use as input a full block of data for each example, recurrent layers break sequence data into time steps, and iterate over them sequentially. This process allows the gradient descent algorithm, known as backpropagation through time, to share parameters across time steps as well as make inferences based on the ordering of SNPs—i.e. to have a memory of allelic associations. The bidirectional attribute of the GRU layer simply means that each example is duplicated and reversed, so the sequence data are analyzed from both directions and then merged by concatenation.

Performance on Simulated Chromosomes

As a proof of principle, we performed coalescent simulations using msprime (***Kelleher et al., 2016***) to generate whole chromosome samples using a fine scale genetic map estimated from *D. melanogaster* (***Comeron et al., 2012***). We then used ReLERNN to estimate the landscape of recombination for these examples. ReLERNN is able to predict the per-base recombination landscape along a simulated chromosome to a high degree of accuracy across a wide range of realistic parameter values, assumptions, and sample sizes ($R^2 \geq 0.82$; Mean absolute error (MAE) $\leq 1.28 \times 10^{-8}$). Importantly, the accuracy of ReLERNN is only modestly diminished when comparing predictions based on 20 samples ($R^2 = 0.93$; $MAE = 3.72 \times 10^{-9}$; ***Figure 2***) to those based on four samples ($R^2 = 0.82$; $MAE = 6.66 \times 10^{-9}$; ***Figure 2–Figure Supplement 1***). While ReLERNN retains accuracy at small sample sizes, it exhibits somewhat greater sensitivity to both the assumed per-base mutation rate and the assumed maximum ratio of ρ to the population mutation parameter, θ—two mandatory assumptions.

To assess the degree of sensitivity to these mutation rate assumptions, we ran ReLERNN on simulations using an assumed per-base mutation rate both 50% greater and 50% less than the simulated (true) mutation rate. In both scenarios, ReLERNN predicts crossover rates that are highly

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

correlated with the simulated rates ($R^2 > 0.91$). However, in both scenarios MAE is inflated but still modest, and the absolute rates of recombination are underpredicted ($R^2 = 0.91$; $MAE = 1.23 \times 10^{-8}$; ***Figure 2–Figure Supplement 2***) and overpredicted ($R^2 = 0.94$; $MAE = 1.28 \times 10^{-8}$; ***Figure 2–Figure Supplement 3***) when assuming a mutation rate less than or greater than the true per-base mutation rate, respectively. Together these results suggest that ReLERNN is in fact learning information about the ratio of crossovers to mutations, and while ReLERNN is highly robust to errant assumptions when predicting relative recombination rates within a genome, caution must be taken when comparing absolute rates between organisms with large differences in per-base mutation rate estimates.

ReLERNN compares favorably to competing methods, especially for small sample sizes and under model misspecification

To assess the accuracy of ReLERNN relative to existing methods, we took a comparative approach whereby we made predictions on the same set of simulated test chromosomes using methods that differ broadly in their approaches. Specifically, we chose to compare ReLERNN against two types of machine learning methods—a boosted regression method, FastEPRR (***Gao et al., 2016***), and a convolutional neural network (CNN) recently described in ***Flagel et al.*** (***2018***)—and LDhat (***McVean et al., 2002***), a widely cited approximate-likelihood method. We independently simulated 10^5 chromosomes using msprime (***Kelleher et al., 2016***) (parameters: $n \in \{4, 8, 16, 32, 64\}$, $priorLowsRho = 0.0$, $priorHighsRho = 5e^{-8} \times 1.25$, $priorLowsMu = 2.5e^{-8} \times 0.75$, $priorHighsMu = 2.5e^{-8} \times 1.25$, $ChromosomeLength = 3e^5$). Half of these were simulated under demographic equilibrium and half were simulated under a realistic demographic model (based on the out-of-Africa expansion of European humans; see Materials and Methods). We show that ReLERNN outperforms all other methods, exhibiting significantly reduced absolute error under both the demographic model and under equilibrium assumptions ($T \leq -31$; $P < 10^{-16}$; *post hoc* Welch's two sample *t*-tests for all comparisons; ***Figure 3***). Importantly, ReLERNN is also more accurate than all methods we compared for each of the tested samples sizes, although all methods generally performed well with larger sample sizes.

We also sought to assess the robustness of ReLERNN to demographic model misspecification, whereby different generative models are used for simulating the training and test sets—e.g. training on assumptions of demographic equilibrium when the test data was generated by a population bottleneck. Methods robust to this type of misspecification are crucial, as the true demographic history of a sample is often unknown and methods used to infer population size histories can disagree or be unreliable (see ***Figure 5–Figure Supplement 1***). Moreover, population size changes alter the landscape of LD across the genome (e.g ***Slatkin, 1994***; ***Rogers, 2014***), and thus have the potential to reduce accuracy or produce biased recombination rate estimates.

To this end, we trained ReLERNN on examples generated under equilibrium and made predictions on 5000 chromosomes generated by the human demographic model specified above (and also carried out the reciprocal experiment). We compared ReLERNN to both the CNN and LDhat, whereby all methods were similarly misspecified (see Materials and Methods). We found that ReLERNN outperforms both the CNN and LDhat, exhibiting significantly lower absolute error under both directions of demographic model misspecification ($T \leq -26$; $P_{WTT} < 10^{-16}$ for all comparisons; ***Figure 4***). Interestingly, we show that the error attributed to model misspecification (termed marginal error; see Materials and Methods) is significantly greater when ReLERNN was trained on equilibrium simulations and tested on demographic simulations than under the reciprocal misspecification ($T = 26.3$; $P_{WTT} < 10^{-16}$; ***Figure 4–Figure Supplement 1***). While this is true, it is important to note that marginal error is quite modest in both directions of misspecification ($< 1.30 \times 10^{-9}$; ***Figure 4–Figure Supplement 1***), suggesting that the additional information gleaned from an informative demographic model is limited.

Differences in the ratio of homologous gene conversion events to crossovers can also bias the inference of recombination rates, as conversion tracts break down LD within the prediction window (***Gay et al., 2007***; ***Przeworski and Wall, 2001***). We treated the effect of gene conversion as

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

another form of model misspecification by training on examples that lacked gene conversion and testing on examples that included gene conversion. As ReLERNN uses msprime for all training simulations, and msprime cannot currently simulate gene conversion, we generated all test set simulations with ms (***Hudson, 2002***). We found that including gene conversion in our simulations biased our predictions, resulting in an overestimate of the true recombination rate (***Figure 4–Figure Supplement 2***). Moreover, the magnitude of this bias increased with the ratio of gene conversion events to crossovers. As expected, we also observed a similar pattern of bias for LDhat, although the magnitude of bias for LDhat was somewhat less than that exhibited by ReLERNN (***Figure 4–Figure Supplement 2***).

Recombination landscapes are largely concordant among populations of African *D. melanogaster*

Using our method, we characterized the genome-wide recombination landscapes of three populations of African *D. melanogaster* (sampled from Cameroon, Rwanda, and Zambia). Each population was derived from the sequencing of 10 haploid embryos (detailed in ***Lack et al., 2015***; ***Pool et al., 2012***), hence these data represent an excellent opportunity to exploit ReLERNN's high accuracy on small sample sizes. We first sought to model the demographic history of each population, as ReLERNN can simulate training data under demographic models inferred by three published software methods—stairwayplot (***Liu and Fu, 2015***), SMC++ (***Terhorst et al., 2016***), and MSMC (***Schiffels and Durbin, 2014***). Using all three methods, we show that inferred historical population sizes are unreliable for these populations—no two methods recapitualte the same history, and the histories generated by MSMC vary dramatically depending on the number of samples used (***Figure 5–Figure Supplement 1***, ***Figure 5–Figure Supplement 2***). For these reasons, and because results from our simulations suggest that marginal error due to demographic misspecification is quite low for our method (above; ***Figure 4–Figure Supplement 1***), we decided to simulate our training data under the assumptions of demographic equilibrium.

Using ReLERNN, we discovered that the fine-scale recombination landscapes are highly correlated among all three populations of *D. melanogaster* (genome-wide mean pairwise Spearman's $\rho = 0.76$; $P < 10^{-16}$; 100 Kb windows; ***Figure 5***). The genome-wide mean pairwise coefficient of determination between populations was somewhat lower, $R^2 = 0.63$ ($P < 10^{-16}$; 100 Kb windows), suggesting there may be important population-specific differences in the fine-scale drivers of allelic association. These differences may also contribute to within-chromosome differences in recombination rate between populations. Indeed, we estimate that mean recombination rates are significantly different among populations for all chromosomes with the exception of chromosome 3L ($P \leq 3.78 \times 10^{-4}$; one-way analysis of variance). Post-hoc pairwise comparisons suggest that this difference is largely driven by an elevated rate of recombination in Zambia, identified on all chromosomes ($P \leq 8.21 \times 10^{-4}$; Tukey's HSD tests) except for 3L ($P_{HSD} \geq 0.15$). ReLERNN predicts the recombination rate in simulated test sets to a high degree of accuracy for all three populations ($R^2 \geq 0.93$; $P < 10^{-16}$; ***Figure 5–Figure Supplement 3***), suggesting that we have sufficient power to discern fine-scale differences in per-base recombination rates across the genome.

When comparing our recombination rate estimates to those derived from experimental crosses of North American *D. melanogaster* (reported in ***Comeron et al., 2012***), we find that the coefficients of determination averaged over all three populations were $R^2 = 0.46, 0.70, 0.47, 0.08, 0.73$ for chromosomes 2L, 2R, 3L, 3R, and X, respectively (***Figure 5–Figure Supplement 4***; 1 Mb windows). These results differ from those observed by ***Chan et al. (2012)***, who compared 22 *D. melanogaster* sampled from the same Rwandan population to the FlyBase map and found $R^2 = 0.55, 0.63, 0.45, 0.42, 0.41$ for the same chromosomes. The minor differences we observed between methods for chromosomes 2L, 2R, and 3L can likely be attributed to the fact that we are comparing estimates from two different methods, using different African flies, to a different experimentally derived map. However, the larger differences found between methods for chromosomes 3R and the X seem less likely attributable to methodological differences. Importantly, African *D. melanogaster* are known to harbor

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-4.0 International license.

large polymorphic inversions (***Corbett-Detig and Hartl, 2012***; ***Lack et al., 2015***), often at appreciable frequencies. For example, the inversion *In(3R)K* segregates in our Cameroon population at $p = 0.9$. It is potentially these differences in inversion frequencies that contribute to the exceptionally weak correlation observed using our method for chromosome 3R and the larger differences between methods for chromosome X.

An important cause of population-specific differences in recombination landscapes might be population-specific differences in the frequencies of chromosomal inversions, as recombination is expected to be strongly suppressed between standard and inversion arrangements. Segregating inversions in *D. melanogaster* have been shown to affect broad patterns of chromosomal variation, and are thought to have quite recent origins when taken together (***Corbett-Detig and Hartl, 2012***). To test for an effect of inversion frequency on our measurement of recombination rates, we resampled haploid genomes from Zambia to create sampled populations with the cosmopolitan inversion *In(2L)t* segregating at varying frequencies, $p \in \{0.0, 0.2, 0.6, 1.0\}$. In Zambia, *In(2L)t* segregates at $p = 0.22$ (***Lack et al., 2015***), suggesting that recombination within the inversion breakpoints may be strongly suppressed in individuals with the inverted arrangement relative to those with the standard arrangement. Moreover, *In(2L)t* arose recently, likely within the past 100,000 years (***Corbett-Detig and Hartl, 2012***). For these reasons, we predict that the inferred recombination rate should decrease as the low-frequency inverted arrangement is increasingly overrepresented in the set of sampled chromosomes (i.e. as more of the samples contain the high-LD inverted arrangements). As predicted, we found a strong effect of the sample frequency of *In(2L)t* on estimated rates of recombination for chromosome 2L in Zambia (***Figure 6***). Recombination rates are negatively correlated with inversion frequency in our sample, not only within the inversion, but also in regions 3 Mb outside the inversion (flanking regions) ($\rho_{Spearman's} = -1$; $P = 0.04$ for both comparisons). We also see a similar negative correlation outside the flanking regions, although this association is weakened relative to that within or flanking the inversion (***Figure 6***). Importantly, varying the size of the flanking regions (from 1-5 Mb) produces patterns that are qualitatively identical, suggesting that the effect of inversions on recombination suppression extends far beyond the inversion breakpoints themselves (***Figure 6–Figure Supplement 1***).

While the effect of inversion frequency on recombination rates may extend beyond the inversion breakpoints, we expect that rates of recombination should be correlated with distance to the inversion breakpoint on smaller spatial scales. To test this we looked at the recombination rates in our African *D. melanogaster* populations, binned by distance to the nearest inversion breakpoints segregating in these populations. Importantly, we curated the samples for our population comparisons by seeking to match the frequency of each inversion segregating in our samples with its true population frequency, as measured in the whole of the DGN database (see Materials and Methods). We show that recombination rates in the flanking regions are positively correlated with distance to inversion breakpoints in both Rwanda and Zambia ($\rho_{Spearman's} = 1$; $P = 0.04$ for both comparisons) but not in Cameroon ($\rho_{Spearman's} = 0.8$; $P = 0.17$; ***Figure 7***). Likewise, recombination rates in the inversion interior (> 2 Mb from the breakpoints) are expected to be higher than in those regions immediately surrounding the breakpoints. However, with the exception of Cameroon (Inversion interior compared to < 250 Kb from breakpoint; $P_{WTT} = 0.035$), we did not observe this pattern ($P_{WTT} \geq 0.057$; ***Figure 7***).

To further explore population-specific differences in recombination landscapes we took a statistical outlier approach, whereby we define two types of recombination rate outliers—global outliers and population-specific outliers (see Materials and Methods). Global outliers are characterized by windows with exceptionally high variance in rates of recombination between all three populations (***Figure 5***; red triangles) while population-specific outliers are those windows where the rate of recombination in one population is strongly differentiated from the rates in the other two populations (***Figure 5***; population-colored triangles). We find that population-specific outliers, but not global outliers, are significantly enriched within inversions ($P = 0.005$; randomization test; ***Figure 5***; grey boxes). Moreover, this enrichment remains significant when extending the inversion boundaries

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-4.0 International license.

by up to 250 Kb ($P_{rand} \leq 0.004$). However, extending the inversion boundaries beyond 250 Kb, or restricting the overlap to windows surrounding only the breakpoints (250 Kb, 500Kb, 1 Mb, 2 Mb), erodes this pattern ($P_{rand} \geq 0.055$ for all comparisons), suggesting that the role for inversions in generating population-specific differences in recombination rates is complex, at least for these populations.

Selection is another important factor that may confound the inference of recombination rates. For instance selective sweeps generate localized patterns of high LD on either side of the sweep site (***Kim and Nielsen, 2004***; ***Schrider et al., 2015***), thus regions flanking selective sweeps may mimic regions of reduced recombination. Inasmuch population-specific selective sweeps are expected to contribute to population-specific differences in recombination rate estimates. We used diploS/HIC (***Kern and Schrider, 2018***) to identify hard and soft selective sweeps in our African *D. melanogaster* populations, and we tested for an excess of recombination rate outliers overlapping with windows classified as sweeps. In total, diploS/HIC classified 27.4%, 28.1%, and 26.8%, of all genomic widows as selective sweeps (either "hard" or "soft") for Cameroon, Rwanda, and Zambia, respectively, when looking at 5kb, non-overlapping windows. The associated False Discovery Rates (FDR) for calling sweeps in these populations were appreciable: 33.9%, 33.1% and 34.7%, respectively (***Figure 5–Figure Supplement 5***). As expected, windows classified as sweeps had significantly lower rates of recombination relative to neutral windows in all three populations ($P_{WTT} \leq 10^{-16}$ for all comparisons; ***Figure 7***). However, we found that neither global nor population-specific outliers were enriched for selective sweeps ($P_{rand} \geq 0.246$ for both comparisons), suggesting that, when treated as a class, recombination rate outliers are not likely driven by sweeps in these populations. When treated separately (i.e. independent permutation tests for each recombination rate outlier window), we identified 7 outliers enriched for sweeps at the $P \leq 0.05$ threshold, corresponding to an expected FDR of 77%. However, given our FDR for calling sweeps in these populations, our measure of the enrichment in overlap with recombination rate outliers is likely to be conservative. Two of these outlier windows may represent potential true positives; an outlier in Cameroon contains 5 out of 6 non-overlapping 5 kb windows classified as "hard" sweeps, the second from Rwanda has 10 out of 12 windows classified as "hard" sweeps ($P_{rand} = 0.0$ for both comparisons). These two recombination rate outlier windows are potentially ripe for future studies on selective sweeps in these populations, and suggest that in at least some instances, selection contributes to observed differences in estimates of recombination rates between *Drosophila* populations.

Discussion

We introduced a new method, ReLERNN, for predicting the genome-wide landscape of per-base recombination rates from phased haplotypes from as few as four samples through the use of deep neural networks. Population genomics, as a field, relies on estimates of recombination rates to understand the effects of diverse phenomena ranging from the impacts of natural selection (***Elyashiv et al., 2016***), to patterns of admixture and introgression (***Price et al., 2009***; ***Brandvain et al., 2014***; ***Schumer et al., 2018***), to polygenic associations in genome-wide association studies (***Bulik-Sullivan et al., 2015***). As befits this need, there has been a long tradition of development of statistical methods for estimating the population recombination parameter, $\rho = 4Nr$ (***Chan et al., 2012***; ***Gao et al., 2016***; ***Hudson and Kaplan, 1985***; ***Hudson, 1987***, ***2002***; ***Li and Stephens, 2003***; ***Lin et al., 2013***; ***McVean et al., 2002***; ***Myers and Griffiths, 2003***; ***Wakeley, 1997***; ***Wall, 2000***; ***Wiuf, 2002***).

We sought to harness the power of deep learning, specifically deep recurrent neural networks, to address the problem of estimating recombination rates, and in so doing, we developed a workflow that reconstructs the genome-wide recombination landscape to a high degree of accuracy from very small sample sizes—e.g. four phased haploid chromosomes. The use of deep learning has recently revolutionized the fields of computer vision (***Krizhevsky et al., 2012***; ***Szegedy et al., 2015***), speech recognition (***Hinton et al., 2012***), and natural language processing (***Sutskever et al., 2014***), and while its use in population genomics has only recently begun, it is anticipated to be similarly fruitful (***Schrider and Kern, 2018***). The natural extension of deep learning to population genomic

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

analyses comes as a result of the ways in which ANNs learn abstract representations of their inputs. In the case of population genomic analyses, the inputs can be naturally represented as DNA sequence alignments, eliminating the need for human oversight (and potentially constraint) in the form of statistical summaries (i.e. compression) of the raw data. ANNs can then learn high-dimensional statistical associations directly from the sequence alignments, and use these to return highly accurate predictions.

ReLERNN utilizes a variant of an ANN, known as a Gated Recurrent Unit (GRU), as its primary technology. GRU networks excel at identifying temporal associations (***Jozefowicz et al., 2015***), and therefore we model our sequence alignment as a bidirectional time series, whereby each ordered SNP represents a new time step along the chromosome. We also model the distance between SNPs using a separate input tensor, and these two inputs are concatenated after passing through the initial layers of the network (see ***Figure 1*** inlay). We demonstrated that ReLERNN can predict a simulated recombination landscape with a high degree of accuracy ($R^2 = 0.93$; ***Figure 2***), and that these predictions remain high, even when using small sample sizes ($R^2 = 0.82$; ***Figure 2–Figure Supplement 1***). Importantly, these predictions compared favorably to those made by a leading composite likelihood method (LDhat; ***McVean et al., 2002***), as well as other machine learning methods (the CNN and FastEPRR; ***Figure 3***). While the abstract nature of the data represented in its internal layers constrains our ability to interpret the exact information ReLERNN relies on to inform its predictions, our experiments using incorrect assumed mutation rates (***Figure 2–Figure Supplement 3***, ***Figure 2–Figure Supplement 2***) suggests that ReLERNN is potentially learning the relative ratio of recombination rates to mutation rates. For these reasons, an extra caveat is warranted—use caution when interpreting the results from ReLERNN as precises measures of the per-base recombination rate unless precise mutation rate estimates are also known.

Demographic model misspecification is another potential source of error that should affect not only deep learning methods targeted at estimating ρ, but also likelihood-based methods. Historical demographic events (e.g. population bottlenecks, rapid expansions, etc.), because they may alter the structure of LD genome-wide, can bias inference of recombination based on genetic variation data. Our simulations demonstrated that while all the methods we tested had elevated error in the context of demographic model misspecification, ReLERNN remained the most accurate across all misspecification scenarios (***Figure 4***). While we caution against generalizing too much from this experiment, the model misspecification tested here was extreme: we are replacing a human-like demography of a bottleneck followed by exponential growth with a model of demographic equilibrium. We suspect that ReLERNN, by using an RNN, is able to encode higher-order allelic associations across the genome, for instance three-locus or four-locus linkage disequilibrium, and in so doing capture more of the information available than traditional methods that use composite likelihoods of two-locus LD summaries. Additionally, there are clear opportunities for future improvements to ReLERNN. For instance, our simulation studies demonstrated that the RNN used by ReLERNN is also sensitive to gene conversion events(***Figure 4–Figure Supplement 2***), thus the joint estimation of rates of recombination and gene conversion may be quite feasible. Ultimately, it remains far from clear what network architectures will be best suited for population genetic inference, though we remain optimistic that ANNs will prove useful for a variety of applications in the field.

A natural application of ReLERNN, due in part to its high accuracy with small sample sizes, was to characterize and compare the recombination landscapes for multiple populations of African *D. melanogaster*, for which few populations with large samples sizes are currently available. Previous estimates of genome-wide fine-scale recombination maps in flies have focused on characterizing recombination in experimental crosses (***Comeron et al., 2012***), or by running LDhat (or the related LDhelmet) on populations with relatively moderate sample sizes (i.e. ≥ 22 samples) (***Chan et al., 2012***; ***Langley et al., 2012***). Here, we applied ReLERNN to three populations for which at least ten haploid embryos were sequenced: Cameroon, Rwanda, and Zambia (***Lack et al., 2015***; ***Pool et al., 2012***). Generally, recombination landscapes were well correlated among populations. Mean pair-

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-4.0 International license.

wise coefficients of determination among all three populations were $R^2 = 0.69, 0.61, 0.77, 0.43, 0.66$ for chromosomes 2L, 2R, 3L, 3R, and X, respectively. These correlations are notably lower than those observed in humans (***Myers et al., 2005***) and mice (***Wang et al., 2017***), and one potential biological cause for this large difference could be the cosmopolitan chromosomal inversions that segregate in African *D. melanogaster* (***Corbett-Detig and Hartl, 2012***; ***Lack et al., 2015***).

We demonstrated a significant negative association between inversion sample frequency and recombination rate as inferred by ReLERNN through experimentally manipulating the frequency of the inversion karyotype in our sample (***Figure 6***). Our results suggest that recombination suppression extends well beyond the predicted breakpoints of the inversion (at least 5 Mb beyond in the case of *In(2L)t*; ***Figure 6–Figure Supplement 1***). This large-scale suppression of recombination due to inversions in *Drosophila* has been observed both directly in experimental crosses (***Dobzhansky and Epling, 1948***; ***Novitski and Braver, 1954***; ***Kulathinal et al., 2009***; ***Miller et al., 2016***; ***Fuller et al., 2018***), and indirectly from patterns of variation surrounding known inversion breakpoints (***Corbett-Detig and Hartl, 2012***; ***Langley et al., 2012***). While it is true that the negative relationship between inversion frequency and recombination should only exist for inversions segregating at low frequencies (e.g. crossover suppression is not expected in inversion homozygotes), we predict a negative relationship to dominate in these populations, as the majority of polymorphic inversions are young, segregate at low frequencies, and show elevated LD along their lengths perhaps due to the actions of natural selection (***Corbett-Detig and Hartl, 2012***; ***Lack et al., 2015***).

While polymorphic inversions exert strong effects on recombination landscapes, support for their role in explaining the most diverged regions among populations was mixed—we found that population-specific recombination rate outliers, but not global outliers, were significantly enriched within the inversions known to segregate in these populations (***Figure 5***). Moreover, our predictions for the relative rates of recombination among populations, based on inversion frequencies per chromosome, were largely not met—the inversions *In(2L)t*, *In(2R)NS*, and *In(3L)Ok* segregate at the highest frequencies in Zambia, yet this population also has the highest average recombination rate for these three chromosomes. Chromosome 3R, however, did match these predictions, having inversions segregating at the highest frequencies of any chromosome (e.g. $p_{In(3R)K} = 0.9$ in Cameroon) and also both the lowest coefficient of determination ($R^2 = 0.43$) and population-specific recombination rates ranked in accordance with inversion frequencies (***Figure 5***).

Interestingly, while we identified two individual outlier regions characterized by numerous selective sweeps, we did not observe a significant enrichment of sweeps overlapping either global or population-specific outliers when these outliers were treated as a class of genomic elements. This is perhaps surprising, given that selective sweeps are known to create characteristic elevations of LD (***Kim and Nielsen, 2004***), and perhaps could mimic regions with very divergent levels of recombination in a population-specific way. A number of other evolutionary forces might explain the existence of our outlier regions as well. For example, mutation rate heterogeneity along the chromosomes could, in principle, generate spurious peaks or troughs in our estimates of recombination rate, as ReLERNN in effect scales its per-base recombination rate estimates by a mutation rate that is assumed to be constant along the chromosome (***Figure 2–Figure Supplement 3***, ***Figure 2–Figure Supplement 2***). Moreover, introgression from diverged populations might affect patterns of allelic association in a a local way along the genome (***Schrider et al., 2018***; ***Schumer et al., 2018***). Taken together, our results suggest that while both inversions and selection can influence population-specific differences in the landscape of recombination, the preponderance of these differences likely have complex causes.

In this report we described ReLERNN, a novel deep learning method for inferring fine-scale rates of recombination across the genome. While ReLERNN currently stands as a functional end-to-end pipeline for measuring recombination rates, the modular design herein presents a number of important opportunities for extension, with the potential to address myriad questions in population genomics. For example, while ReLERNN is currently designed to use phased haplotype data as its input, we see no reason why unphased, diploid genotypes couldn't be substituted (e.g. ***Flagel***

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

et al., 2018). Moreover, the RNN structure we exploit here could be used for inference of the distribution of selection coefficients and/or migration rates from natural populations. In addition, ReLERNN presents an excellent opportunity for the implementation of transfer learning, whereby ReLERNN could be trained in-house on an otherwise prohibitively extensive parameter space, allowing end-users to make accurate predictions by generating only a small fraction of the current number of simulations and training epochs presently required. The application of machine learning, and deep learning in particular, to questions in population genomics is ripe with opportunity. ReLERNN provides a platform for jumping off, that we hope to see advance our understanding of both population genetics and adaptation itself.

Materials and Methods

The ReLERNN workflow

Here we briefly describe ReLERNN, a software package for accurately estimating a genome-wide recombination landscape from as few as four phased chromosomes. The ReLERNN workflow proceeds by the use of four python modules—ReLERNN_SIMULATE, ReLERNN_TRAIN, ReLERNN_PREDICT, and ReLERNN_BSCORRECT (***Figure 1***). The first three modules are mandatory, and include functions to calculate Watterson's estimator and historical population sizes, functions for simulating the training set, functions for training the neural network, and functions for reporting rates of recombination along the chromosomes. The fourth module, ReLERNN_BSCORRECT, is optional (though recommended) and includes functions for estimating 95% confidence intervals and implements a correction function to reduce biases that may arise during training. The output from ReLERNN is a list of genomic windows and their corresponding recombination rate prediction (reported as per-base crossover events), along with 95% confidence intervals if the optional ReLERNN_BSCORRECT module was used.

Parameter estimation and coalescent simulation

ReLERNN takes as input a VCF file of phased biallelic variants, which can either be coded as nucleotides or ancestral/derived states (i.e. 0/1). A minimum of four sample chromosomes must be included, and users should ensure proper filtering of the input file beforehand—e.g. excluding low-coverage or low-quality sites, non-biallelic sites, and missing data. ReLERNN also requires the user to provide an assumed per-base mutation rate and an assumed maximum value for the ratio ρ/θ. These parameters are used to set an acceptable window size for prediction, by restricting the total number of segregating sites in each window to remain below a critical threshold. ReLERNN therefore uses a dynamic window size to reduce the probability of training failure due to having too many, or too few, segregating sites present in a window (e.g. experimental trials showed that the training loss function eventually returns NaNs when training on windows containing multiple thousands of segregating sites). As a result, the output predictions file may return different window sizes for different chromosomes, even within the same genome. For comparing rates between populations, an optional script ("force_window_size_predictions.py") is provided to force rates to conform to a given window size. This is accomplished by taking a weighted average of recombination rates, whereby rates are weighted by the fraction of overlap between their original window positions and the new forced window positions.

Once the appropriate window sizes have been estimated, ReLERNN_SIMULATE uses the coalescent simulation software, msprime (***Kelleher et al., 2016***), to independently generate 10^5 training examples and 10^3 validation and test examples. By default, these simulations are generated under assumptions of demographic equilibrium, using a range of per-base mutation and recombination rates. However, ReLERNN can optionally simulate under a demographic history inferred by one of three programs: stairwayplot (***Liu and Fu, 2015***), SMC++ (***Terhorst et al., 2016***), or MSMC (***Schiffels and Durbin, 2014***), and the handling of output from these programs is fully integrated into ReLERNN_SIMULATE. This provides users the ability to model a demographic history and to estimate rates of recombination from different files (e.g. one that includes only intergenic sites). When each

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

simulation is completed, ReLERNN dumps both the genotype matrix and a vector of the positions for every SNP into a temporary .npy file.

Sequence batch generation and network architecture

To reduce the large memory utilization common to the analysis of genomic sequence data, we took a batch generation approach, whereby only small batches of simulations are called into memory at any one time. Data normalization and padding occurs when a training batch is called, by which the genotype and position arrays are read into memory. In ReLERNN, ancestral states are coded as -1, derived states are coded as 1, and both genotype and positions arrays are padded with 0s to the maximum number of segregating sites generated across all examples. In addition, a framing pad of five 0s is applied to both arrays, and the order of samples in each batch is randomly shuffled. The targets for each training batch are the per-base recombination rates used by msprime when simulating each example. These targets are z-score normalized across all training examples. The normalized and padded genotype and position arrays form the input tensors for our neural network.

ReLERNN trains a recurrent neural network with Keras (***Chollet et al., 2015***) using a Tensorflow backend (***Abadi et al., 2015***). The complete details of our neural architecture can be found in the python module "ReLERNN_networks.py", and a simplified flow diagram showing the connectivity between layers can be found in ***Figure 1***. Briefly, the ReLERNN neural network utilizes distinct input layers for the genotype and position tensors, which are later merged using a concatenation layer in Keras. The genotype tensor is first fed to a GRU layer, as implemented with the bidirectional wrapper in Keras, and the output of this layer is passed to a dense layer followed by a dropout layer. On the positions side of the network, the input positions tensor is fed directly to a dense layer and then to a dropout layer. Dropout was used extensively in our network, as hypertuning trials (below) demonstrated significantly improved accuracy when employing dropout relative to networks without dropout. Once concatenated, output from the dropout layer is passed to a final round of dense and dropout layers, and the final dense layer returns a single z-score normalized prediction for each example, which is unnormalized back to units of crossovers per-base. ReLERNN completes 250 training epochs and implements this training using the "Adam" optimizer and a Mean Squared Error (MSE) loss function. Though the number of epochs is user-selectable, the vast majority of networks are sufficiently trained within 250 epochs, largely due to how ReLERNN handles the input tensor size and simulation parameters. Our hyper-tuning trials were completed via a grid search over the set of parameters: Recurrent layer output dimensions (64, 82, 128), Loss function (MSE, MAE), Input merge strategy (concatenate, average), and dense layer dimensionality (64, 128), optimizing for MSE.

Parametric bootstrap analysis and prediction corrections

ReLERNN includes the option to both generate confidence intervals around each predicted recombination rate and to correct for potential biases generated during training using a parametric bootstrapping approach. After the network has been trained and predictions have been generated, users can run ReLERNN_BSCORRECT, which resimulates 10^3 test examples for each of 100 recombination rate bins drawn from the distribution of recombination rates used to simulate the original training set. Predictions are then generated for these 10^5 simulated test examples using the previously trained network, generating a distribution of predictions for each respective recombination rate bin. 95% confidence intervals are calculated from by taking the upper and lower 2.5% rate predictions from this distributions.

The distribution of test predictions can be biased in systematic ways, such as predictably underestimating rates of recombination for those examples with the highest simulated crossover events (***Figure 1–Figure Supplement 1***). These biases may potentially be caused an inability to resolve very high recombination rates with a limited number of informative SNPs. ReLERNN_BSCORRECT, estimates the magnitude of this bias through bootstrapping, and applies a bias correction function to

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-4.0 International license.

the empirical predictions. The bias correction function takes each empirical prediction and identifies the nearest median value in the bootstrap distribution. The correction function then adds to this prediction the difference between this median value and the true recombination rate used to simulate the distribution of test examples at that recombination rate bin. This correction method has the effect of elevating the empirical prediction in regions of parameter space where we are reasonably confident that we are underestimating recombination rates and lowering the prediction in areas where we are likely to be overestimating recombination rates. ReLERNN_BSCORRECT is provided as an optional module in ReLERNN, as the resimulation of 10^5 test examples is computationally expensive and may not be warranted in all circumstances.

Testing the accuracy of ReLERNN on simulated recombination landscapes

To test the accuracy of ReLERNN at recapitulating a dynamic recombination landscape, we ran our complete ReLERNN workflow on simulation data replicating chromosome 2L of *D. melanogaster*. Using crossover rates estimated by **Comeron et al.** (**2012**), we simulated varying numbers of samples of *D. melanogaster* chromosome 2L with msprime using the RecombinationMap class. Simulated samples were exported to a VCF file using ploidy = 1, and all simulations were generated under demographic equilibrium. We used these simulated VCF files as the input to our ReLERNN pipeline, and ran all ReLERNN modules with default parameters, with the exception of varying the assumed per-base mutation rate and the assumed maximum ratio of ρ to θ. Assumed mutation rates were varied from 50% less than the rate used in simulations (true rate) to 50% greater than the true rate. Likewise, the ratio of ρ to θ was either held constant, resulting in the training set containing on average higher or lower per-base recombination rates than the true rate, or was adjusted to correctly reflect the true maximum per-base recombination rate used—i.e. approximately 1.2×10^{-7} crossovers per base.

Comparative methods

We chose to compare ReLERNN to three published methods for estimating recombination rates—FastEPRR (**Gao et al., 2016**), a 1-dimensional CNN recently described in **Flagel et al.** (**2018**) and LDhat (**McVean et al., 2002**). We generated a training set (used by ReLERNN and the CNN) with 10^5 examples and tested each method on an identical set of 5×10^3 simulation examples for testing. We generated two classes of simulations, one simulated under demographic equilibrium and one using a demographic history derived from European humans (CEU model; detailed in "ReLERNN_demographic_models.py"; **Tennessen et al., 2012**; **Gravel et al., 2011**). Both classes of simulations were generated for $n \in \{4, 8, 16, 32, 64\}$, where n is the number of chromosomes sampled from the population. All simulations were generated in msprime with the common set of parameters: $priorLowsRho = 0.0$, $priorHighsRho = 5e^{-8} \times 1.25$, $priorLowsMu = 2.5e^{-8} \times 0.75$, $priorHighsMu = 2.5e^{-8} \times 1.25$, $ChromosomeLength = 3e^5$, whereby values for both per-base mutation and recombination rates were drawn from a uniform distribution between the low and high priors.

For both ReLERNN and the CNN, the same training set consisting of 10^5 examples was used to train each neural network, and the same test examples were used to compare the predictions produced by each method. Comparisons with LDhat where made using the above training examples to parameterize the generation of independent coalescent likelihood lookup tables. For each set of examples of sample size N, we calculated the maximum value of ρ from the training set and the average per-base values for θ for the test examples, using Watterson's estimator. These parameter values were given to LDhat's $complete$ function for the lookup table generation, and the resulting table was used to make predictions on our 5×10^3 test examples using the $pairwise$ function. Comparisons with FastEPRR were made by transforming the genotype matrices resulting from our test simulations into fasta-formated input files, and running the FastEPRR_ALN funtion (using format = 1) in R. As both LDhat and FastEPRR predict ρ, the resulting predictions were transformed to per-base recombination rates for comparison with ReLERNN using the function $r = \frac{\rho_{pred} \times \mu_{true}}{\theta_W}$, whereby ρ_{pred} is the prediction output by each method, and θ_W and μ_{true} are Watterson's estimator

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

and the true per-base mutation rate used in the simulation example, respectively. To compare accuracy among methods we directly compared the distribution of absolute errors ($|r_{predicted} - r_{true}|$) for each method for each set of examples of sample size N.

To test the effects of model misspecification on predictions, we simply directed ReLERNN and the CNN to use a training set generated under demographic equilibrium for making predictions on a test set generated under the CEU model, and vice versa. To test for the effects of model misspecification in LDhat, we generated a lookup table using parameter values estimated from the misspecified training set (e.g. the lookup table used for predicting the CEU model test set was generated by using parameter values directly inferred from training simulations under equilibrium. We did not directly test the effect of model misspecification using FastEPRR, as this method takes as input only a fasta sequence file, and therefore the internal training of the model was not able to be separated from the input sequences. To address the effects of model misspecification, we also directly compared the distribution of absolute errors ($|r_{predicted} - r_{true}|$). Additionally, we compared the marginal error directly attributable to model misspecification among methods. We defined marginal error as $\epsilon_m - \epsilon_c$, where ϵ_m and ϵ_c are equal to $|r_{predicted} - r_{true}|$ when the model is misspecified and correctly specified, respectively. We simulated gene conversion test sets using ms (***Hudson, 2002***), with a mean conversion tract length of 352 bp (corresponding to the mean empirically derived tract length in *D. melanogaster* (***Hilliker et al., 1994***)) and simulated a ratio of conversion events to crossover events of $0, 1, 2, 4$, and 8.

Recombination rate variation in *D. melanogaster*

We obtained *D. melanogaster* population sequence data from the *Drosphila* Genome Nexus (DGN; https://www.johnpool.net/genomes.html; ***Lack et al., 2015***; ***Pool et al., 2012***). We converted DGN "consensus sequence files" to VCF format using custom python scripts, excluding all non-biallelic sites and sites containing missing data. We chose to analyze populations from Cameroon, Rwanda, and Zambia, as these populations contained at least 10 haploid embryo sequences per population and each population included multiple segregating chromosomal inversions (supplemental table 1). To ensure roughly equivalent power to compare rates among populations, we downsampled both Rwanda and Zambia to 10 chromosomes. We selected individual haploid genomes for each population by requiring that our sampled inversion frequencies for each of the six segregating inversions—*In(1)Be*, *In(2L)t*, *In(2R)NS*, *In(3L)Ok*, *In(3R)K*, and *In(3R)P*—closely approximate their population frequencies as measured in the complete set of haploid genomes for that population. All sample accessions and their corresponding inversion frequencies are located in the supporting materials.

Before running ReLERNN, we first set out to model the demographic history for each population using each of three methods: stairwayplot (***Liu and Fu, 2015***), SMC++ (***Terhorst et al., 2016***), and MSMC (***Schiffels and Durbin, 2014***). With the exception of MSMC, all methods were run using default parameters. For MSMC, the use of default parameters generated predictions that were unusable (***Figure 5–Figure Supplement 2***). For these reasons, and after direct communication with MSMC's authors, we determined that running MSMC with a sample size of two chromosomes would be the most appropriate. Ultimately we decided to run our ReLERNN pipeline with simulations generated under demographic equilibrium [options: –estimateDemography False –assumedMu 3.27e-9 –upperRhoThetaRatio 35], as estimates of historical population size were unreliable for these data—all three methods produced significantly different demographic histories (***Figure 5–Figure Supplement 1***)—and tests on simulated data suggest little effect of demographic model misspecification (***Figure 4–Figure Supplement 1***). All code required to run our ReLERNN analysis is deposited on GitHub (https://github.com/kern-lab/ReLERNN and https://github.com/kern-lab/ReLERNN-analysis).

We measured the correlation in recombination rates between each African *D. melanogaster* populations in 100 kb sliding windows, as ReLERNN will predict the rates of recombination in slightly different window sizes, depending on θ for each chromosome. The recombination rate for each

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

sliding window was calculated by taking the average of all rate windows predicted by ReLERNN, weighted by the fraction that each window overlapped the larger sliding window. Recombination rate outliers were identified in two ways: as global outliers and population-specific outliers. Global outliers were identified by first calculating the mean and standard deviation in recombination rates for all three populations in each 100 kb sliding window. We then used the top 1% of outliers from the distribution of residuals, after fitting a linear model to the standard deviation on the mean. Population-specific outliers were identified by using a modification of the population branch statistic (herein PBS*; ***Yi et al., 2010***), whereby we replaced pairwise F_{ST} with the pairwise differences in recombination rates. We then used the top 1% of all PBS* scores as our population-specific outliers, with each outlier corresponding to a PBS* score for a single population.

To test the effect of inversion frequency on predicted recombination rates, we resampled 10 haploid chromosomes from the available set of haploid genomes from Zambia to generate sampled populations containing *In(2L)t* at varying frequencies, $p \in \{0.0, 0.2, 0.6, 1.0\}$. We then ran ReLERNN on chromosome 2L for each of these resampled Zambian populations. We classified recombination windows by their overlap with the coordinates of In(2L)t (as defined in ***Corbett-Detig and Hartl, 2012***), defining windows within the breakpoints (inside), windows up to 3 Mb outside the breakpoints (flanking), and windows > 3 Mb outside the breakpoints (outside).

To test the effect of genome-wide inversion breakpoints on differences in recombination landscapes between populations, we classified windows by their overlap with inversion interiors (> 2 Mb inside the inversion breakpoints) and their overlap with windows within 200 Kb, 500 Kb, 1 Mb, and 2 Mb of inversion breakpoints. We tested for an enrichment of both global and population-specific outliers within inversions by randomization tests, whereby we permuted the labels for outliers 10^4 times and counted the overlap with inversions for each permutation to calculate the empirical p-values. We also tested for an effect of selection on recombination rates in these populations, by running diploS/HIC (***Kern and Schrider, 2018***) to detect selective sweeps. We ran diploS/HIC on each population, training on simulations generated under demographic equilibrium. For each population we simulated 2000 training examples from each of the five classes of regions required by diploS/HIC using the coalescent simulation software discoal (***Kern and Schrider, 2016***). For simulations which included sweeps we drew the selection coefficient from a uniform distribution such that $s \sim U(0.0001, 0.005)$, the time of completion of the sweep from $\tau \sim U(0, 0.05)$, and the frequency at which a soft sweep first comes under selection as $f \sim U(0, 0.1)$. We drew θ from $U(65, 654)$ and we drew ρ from an exponential distribution with mean 1799 and the upper bound truncated at triple the mean. For the discoal simulations we simulated 605 kb of data with the goal of classification of the central most 55 kb window. We looked at the overlap with "sweep" windows (those classified as either "hard" or "soft") and those windows classified as "neutral" by diploS/HIC. Our complete diploS/HIC pipeline for these samples is available in the supporting materials online. All statistical tests were completed in R (***R Core Team, 2018***), with the exception of empirical randomization tests, which were completed using Python.

Data availability

ReLERNN is currently available at https://github.com/kern-lab/ReLERNN. Supporting information, tables, and figures will be deposited online at the publication journal.

Acknowledgments

The authors would like to gratefully acknowledge Matthew Hahn, Dan Schrider, and Peter Ralph for their helpful comments and suggestions. This work benefited from access to the University of Oregon high performance computer, Talapas. JRA, JGG, and ADK were supported by NIH award R01GM117241 to ADK. We would also like to thank the Hearth for their fine coffee.

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

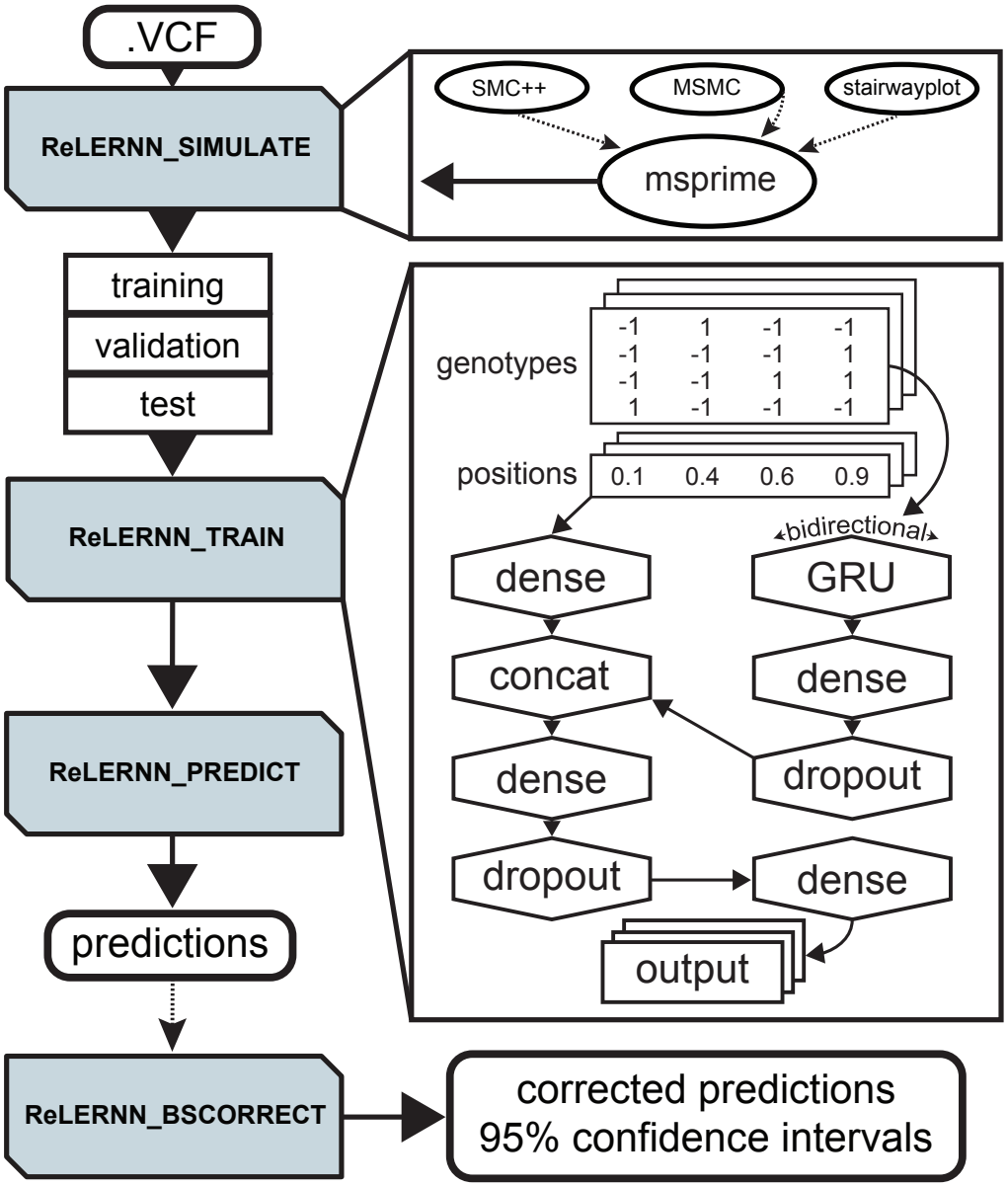

Figure 1. Diagram depicting a typical workflow using ReLERNN's four modules (shaded boxes). ReLERNN_SIMULATE can optionally (dotted lines) utilize output from stairwayplot, SMC++, MSMC to simulate under a demographic history in msprime. The breakout of ReLERNN_TRAIN depicts the GRU network architecture used for training. The input genotype matrix shows alleles encoded as ancestral (-1), derived (1), or padded (0; *not shown*), and the input position matrix shows variant position coded along the real number line (0-1).

Figure 1–Figure supplement 1. Parametric bootstraping results as implemented by ReLERNN. Lines represent the minimum (blue), lower 5% (orange), lower 25% (green), median (red), upper 25% (purple), upper 95% (brown), and maximum (pink) bounds for each of 1000 replicate simulations and predictions (y-axis) across 100 recombination rate bins (x-axis)

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

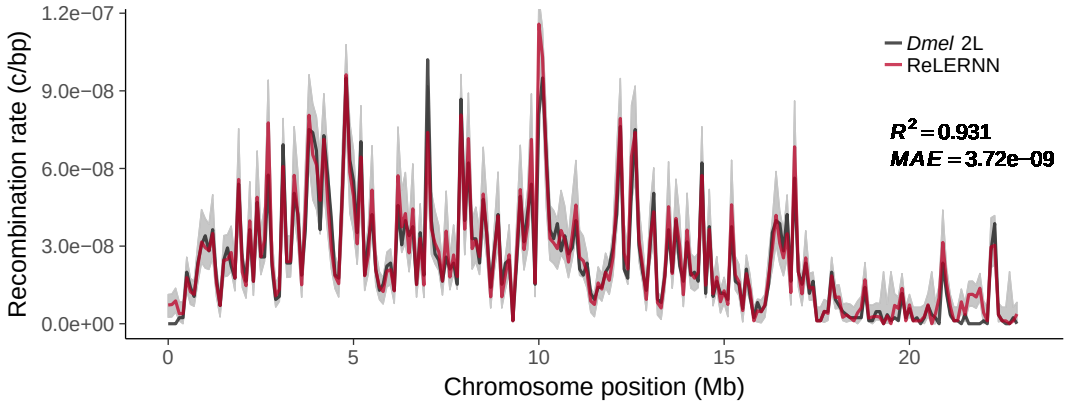

Figure 2. Recombination rate predictions for a simulated *Drosophila* chromosome (black line) using ReLERNN (red line). The recombination landscape was simulated for $n = 20$ chromosomes under mutation-drift equilibrium using msprime (***Kelleher et al., 2016***), with per-base crossover rates derived from *D. melanogaster* chromosome 2L (***Comeron et al., 2012***). Gray ribbons represent 95% confidence intervals. R^2 is reported for the general linear model of predicted rates on true rates and mean absolute error was calculated across all 100 kb windows.

Figure 2–Figure supplement 1. Recombination rate predictions for a simulated *Drosophila* chromosome (black line) using ReLERNN (red line). The recombination landscape was simulated for $n = 4$ chromosomes under mutation-drift equilibrium using msprime (***Kelleher et al., 2016***), with per-base crossover rates derived from *D. melanogaster* chromosome 2L (***Comeron et al., 2012***). Gray ribbons represent 95% confidence intervals. R^2 is reported for the general linear model of predicted rates on true rates and mean absolute error was calculated across all 100 kb windows.

Figure 2–Figure supplement 2. Recombination rate predictions for a simulated *Drosophila* chromosome (black line) using ReLERNN (red line). The recombination landscape was simulated for $n = 20$ chromosomes under mutation-drift equilibrium using msprime (***Kelleher et al., 2016***), with per-base crossover rates derived from *D. melanogaster* chromosome 2L (***Comeron et al., 2012***). Here the per-base mutation rate was assumed to be 50% less than the rate used for simulation. Gray ribbons represent 95% confidence intervals. R^2 is reported for the general linear model of predicted rates on true rates and mean absolute error was calculated across all 100 kb windows.

Figure 2–Figure supplement 3. Recombination rate predictions for a simulated *Drosophila* chromosome (black line) using ReLERNN (red line). The recombination landscape was simulated for $n = 20$ chromosomes under mutation-drift equilibrium using msprime (***Kelleher et al., 2016***), with per-base crossover rates derived from *D. melanogaster* chromosome 2L (***Comeron et al., 2012***). Here the per-base mutation rate was assumed to be 50% greater than the rate used for simulation. Gray ribbons represent 95% confidence intervals. R^2 is reported for the general linear model of predicted rates on true rates and mean absolute error was calculated across all 100 kb windows.

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

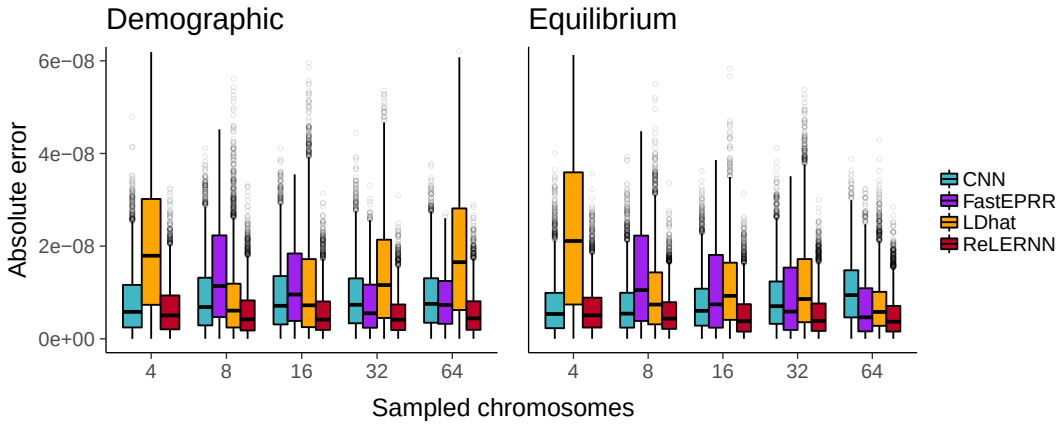

Figure 3. Distribution of absolute errors ($|r_{predicted} - r_{true}|$) for each method across 5000 simulated chromosomes (1000 for FastEPRR). Independent simulations were run under a known demographic history (left) or an assumption of demographic equilibrium (right). Sampled chromosomes indicate the number of independent sequences that were sampled from each msprime (***Kelleher et al., 2016***) coalescent simulation.

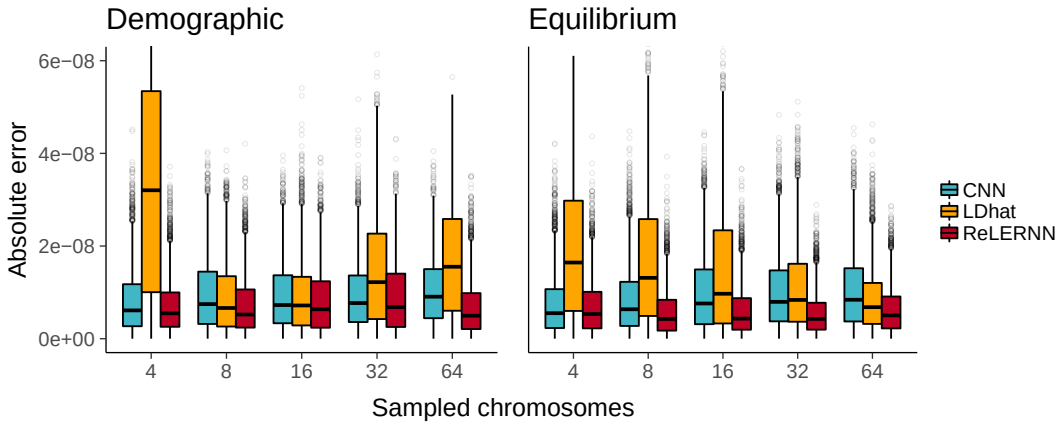

Figure 4. Distribution of absolute errors ($|r_{predicted} - r_{true}|$) for each method across 5000 simulated chromosomes after model misspecification. For the CNN and ReLERNN, predictions were made by training on equilibrium simulations and testing on sequences simulated under a demographic model (left) or training on demographic simulations and testing on sequences simulated under equilibrium (right). For LDhat, the lookup table was generated using parameters values that were estimated from simulations where the model was misspecified in the same way as described for the CNN and ReLERNN above. Sampled chromosomes indicate the number of independent sequences that were sampled from each msprime (***Kelleher et al., 2016***) coalescent simulation.

Figure 4–Figure supplement 1. Distribution of marginal errors attributed to model misspecification across 5000 simulated chromosomes. Predictions were made by training on equilibrium simulations and testing on sequences simulated under a demographic model (left) or training on demographic simulations and testing on sequences simulated under equilibrium (right). Here, marginal errors are represented as $\epsilon_m - \epsilon_c$, where ϵ_m and ϵ_c are equal to $|r_{predicted} - r_{true}|$ when the model is misspecified and correctly specified, respectively. Sampled chromosomes indicate the number of independent sequences that were sampled from each msprime (***Kelleher et al., 2016***) coalescent simulation.

Figure 4–Figure supplement 2. Distribution of predicted rates of recombination over true rates for 5000 examples simulated with gene conversion and $n = 8$. The ratio of gene conversion to crossovers was drawn from $U(0, c)$, with $c \in \{0, 1, 2, 4, 8\}$. Gene conversion tract lengths were fixed at 352 bp, and all simulations were completed in ms (***Hudson, 2002***).

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

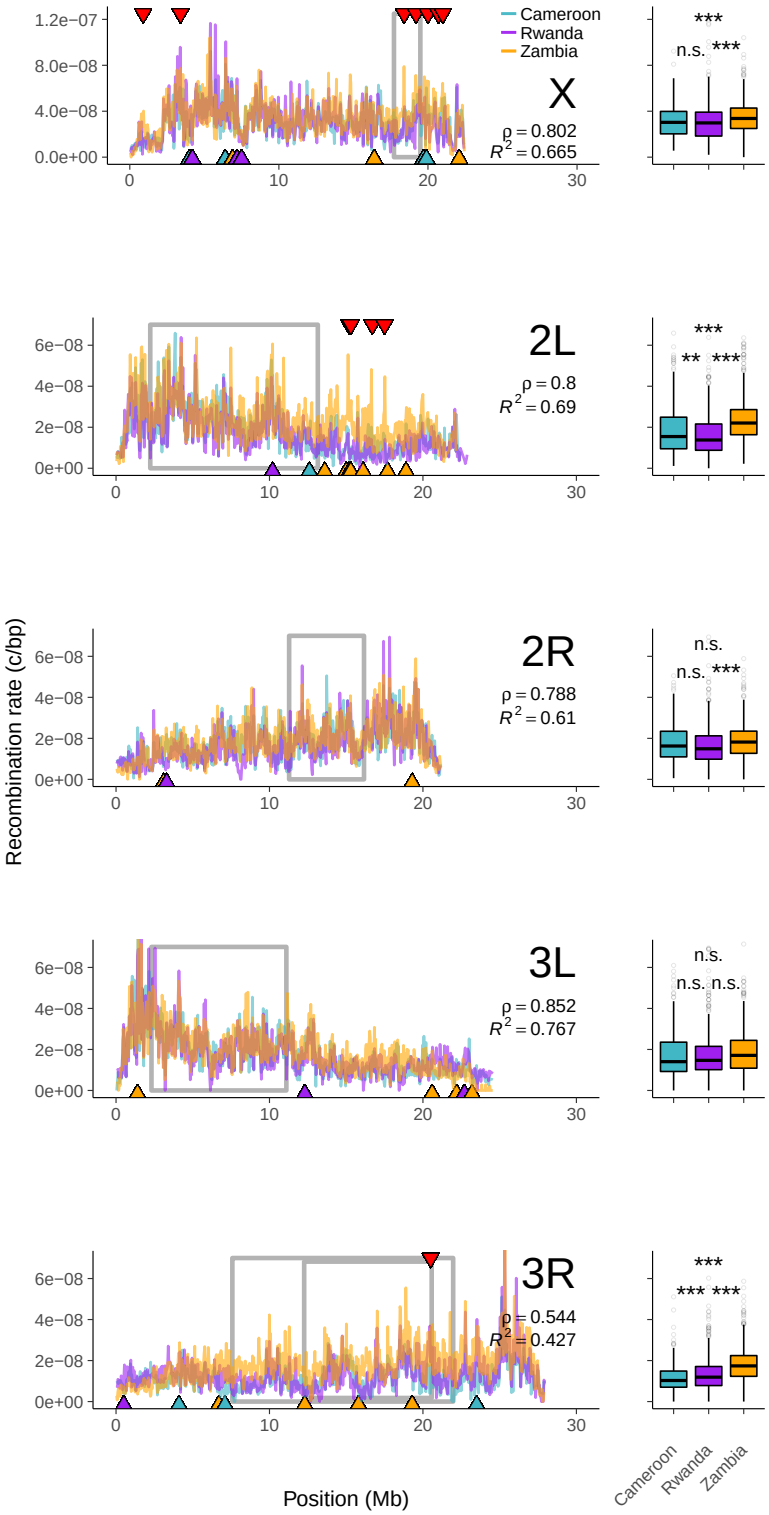

Figure 5. (Left) Genome-wide recombination landscapes for *D. melanogaster* populations from Cameroon (teal lines), Rwanda (purple lines), and Zambia (orange lines). Grey boxes denote the inversion boundaries predicted to be segregating in these samples (***Pool et al., 2012***; ***Corbett-Detig and Hartl, 2012***). Red triangles mark the top 1% of global outlier windows for recombination rate. Blue, purple, and orange triangles mark the top 1% of population-specific outlier windows for recombination rate, with triangle color indicating the outlier population (see Materials and Methods). (Right) Per-chromosome recombination rates for each population. Spearman's ρ and R^2 are reported as the mean of pairwise estimates between populations for each chromosome. $^{**}P < 0.01$ and $^{***}P < 0.001$ are based on Tukey HSD tests for all pairwise comparisons.

Figure 5–Figure supplement 1. Historical population size estimates were inferred for Cameroon, Rwanda, and Zambia using three separate methods, all of which disagree with one another. Inferences are based on 10 samples for both stairwayplot (grey line) and SMC++ (orange line), and 2 samples for MSMC (purple line).

Figure 5–Figure supplement 2. Historical population size estimates were inferred for Cameroon, Rwanda, and

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

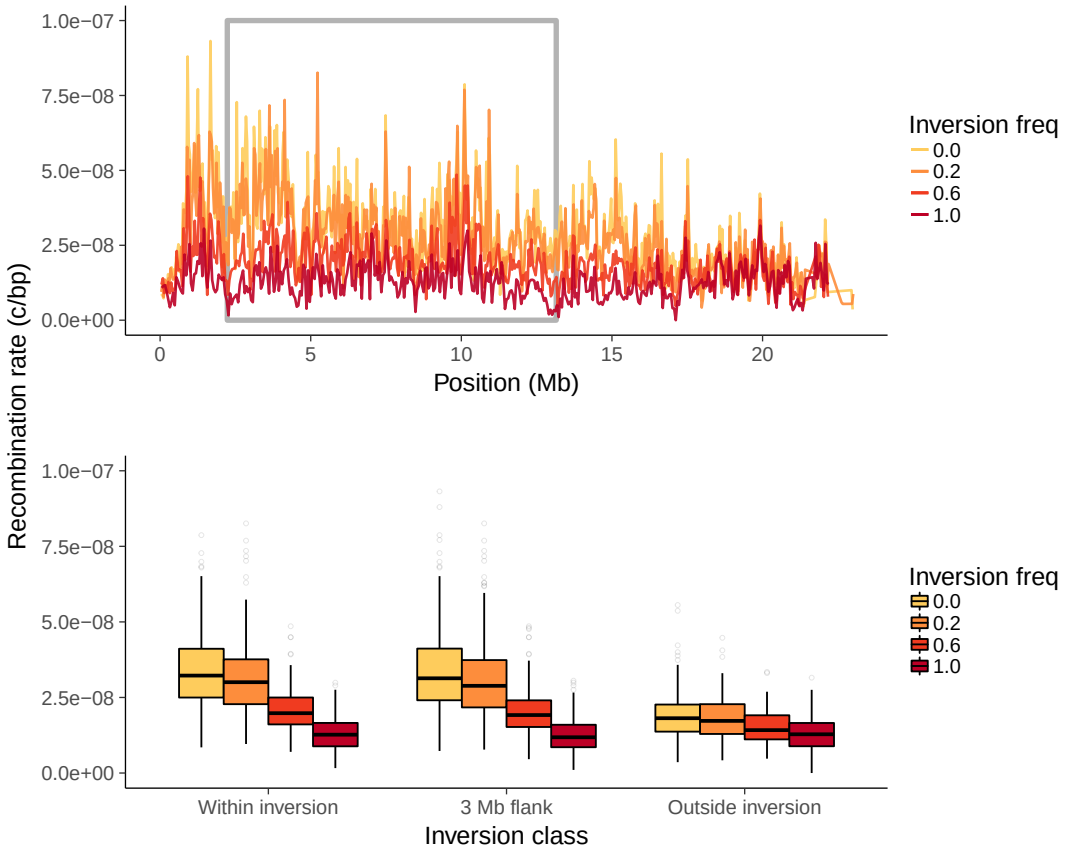

Figure 6. (Top) Recombination landscapes for Zambian *D. melanogaster* surrounding *In(2L)t*, sampled at different inversion frequencies. The grey box denotes the inversion boundaries of *In(2L)t* in *Drosophila* (***Corbett-Detig and Hartl, 2012***). (Bottom) Recombination rate estimates from genomic windows within the inversion, within a 3 Mb region flanking the inversion, and 3 Mb outside the inversion, sampled at different inversion frequencies.

Figure 6–Figure supplement 1. Recombination rate estimates using flanking window sizes from 1-5 Mb. Rates are shown for genomic windows within the inversion, within regions flanking the inversion, and for regions outside both the inversion and flanking regions. All estimates are from chromosome 2L with *In(2L)t* sampled at different inversion frequencies

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-4.0 International license.

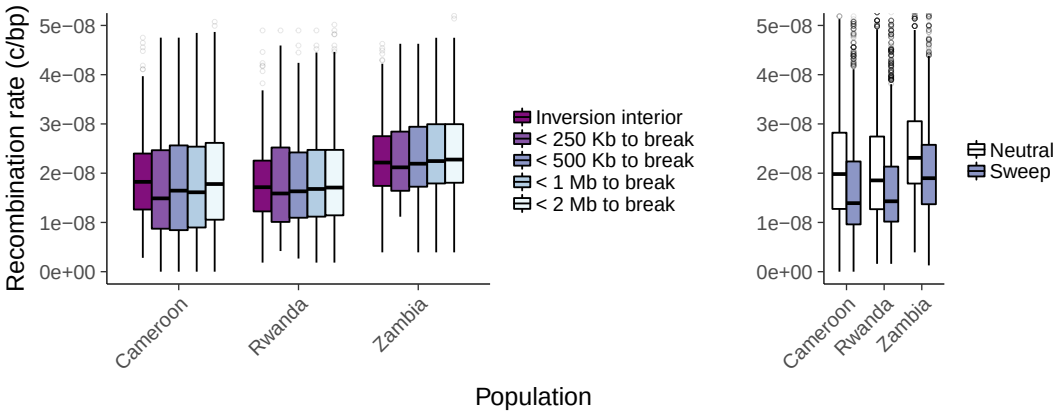

Figure 7. (Left) Recombination rate estimates for genomic windows > 2 Mb inside, < 250 kb surrounding, < 500 kb surrounding, < 1 Mb surrounding, and < 2 Mb surrounding all inversion breakpoints. (Right) Recombination rate estimates for all genomic windows overlapping windows predicted as either hard/soft sweeps (purple) or as neutral (white) by diploS/HIC (***Kern and Schrider, 2018***).

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, et al., TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems; 2015. https://www.tensorflow.org/, software available from tensorflow.org.

Ayala D, Guerrero RF, Kirkpatrick M. Reproductive isolation and local adaptation quantified for a chromosome inversion in a malaria mosquito. Evolution: International Journal of Organic Evolution. 2013; 67(4):946–958.

Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. Speciation and introgression between Mimulus nasutus and Mimulus guttatus. PLoS genetics. 2014; 10(6):e1004410.

Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM, of the Psychiatric Genomics Consortium SWG, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics. 2015; 47(3):291.

Burt A. Perspective: sex, recombination, and the efficacy of selection—was Weismann right? Evolution. 2000; 54(2):337–351.

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

Chan J, Perrone V, Spence JP, Jenkins PA, Mathieson S, Song YS. A Likelihood-Free Inference Framework for Population Genetic Data using Exchangeable Neural Networks. bioRxiv. 2018; https://www.biorxiv.org/content/early/2018/11/05/267211, doi: 10.1101/267211.

Chollet F, et al., Keras. GitHub; 2015. https://github.com/fchollet/keras.

Comeron JM, Ratnappan R, Bailin S. The Many Landscapes of Recombination in Drosophila melanogaster. PLOS Genetics. 2012 10; 8(10):1–21. https://doi.org/10.1371/journal.pgen.1002905, doi: 10.1371/journal.pgen.1002905.

Corbett-Detig RB, Hartl DL. Population Genomics of Inversion Polymorphisms in Drosophila melanogaster. PLOS Genetics. 2012 12; 8(12):1–15. https://doi.org/10.1371/journal.pgen.1003056, doi: 10.1371/journal.pgen.1003056.

Dobzhansky T, Epling C. The suppression of crossing over in inversion heterozygotes of Drosophila pseudoobscura. Proceedings of the National Academy of Sciences of the United States of America. 1948; 34(4):137.

Elyashiv E, Sattath S, Hu TT, Strutsovsky A, McVicker G, Andolfatto P, Coop G, Sella G. A genomic map of the effects of linked selection in Drosophila. PLoS genetics. 2016; 12(8):e1006130.

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

717 **Felsenstein J**. The evolutionary advantage of recombination. Genetics. 1974; 78(2):737–756.

718 **Flagel L**, Brandvain Y, Schrider DR. The Unreasonable Effectiveness of Convolutional Neural Networks in
719 Population Genetic Inference. Molecular Biology and Evolution. 2018 12; 36(2):220–238. https://dx.doi.org/10.
720 1093/molbev/msy224, doi: 10.1093/molbev/msy224.

721 **Fuller ZL**, Koury SA, Leonard CJ, Young RE, Ikegami K, Westlake J, Richards S, Schaeffer SW, Phadnis N. Extensive
722 recombination suppression and chromosome-wide differentiation of a segregation distorter in Drosophila.
723 bioRxiv. 2018; https://www.biorxiv.org/content/early/2018/12/21/504126, doi: 10.1101/504126.

724 **Gao F**, Ming C, Hu W, Li H. New software for the fast estimation of population recombination rates (FastEPRR) in
725 the genomic era. G3: Genes, Genomes, Genetics. 2016; 6(6):1563–1571.

726 **Gay J**, Myers S, McVean G. Estimating meiotic gene conversion rates from population genetic data. Genetics.
727 2007; 177(2):881–894.

728 **Gravel S**, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, , Bustamante CD. Demographic
729 history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences.
730 2011; 108(29):11983–11988. https://www.pnas.org/content/108/29/11983, doi: 10.1073/pnas.1019276108.

731 **Graves A**, Jaitly N, Mohamed A. Hybrid speech recognition with Deep Bidirectional LSTM. In: *2013 IEEE Workshop*
732 *on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*; 2013. p.
733 273–278. https://doi.org/10.1109/ASRU.2013.6707742, doi: 10.1109/ASRU.2013.6707742.

734 **Haenel Q**, Laurentino TG, Roesti M, Berner D. Meta-analysis of chromosome-scale crossover rate variation
735 in eukaryotes and its significance to evolutionary genomics. Molecular Ecology. 2018; 27(11):2477–2497.
736 https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.14699, doi: 10.1111/mec.14699.

737 **Hahn MW**. Molecular population genetics. Sinauer Associates; 2018.

738 **Hartfield M**, Otto SP. Recombination and hitchhiking of deleterious alleles. Evolution: International Journal of
739 Organic Evolution. 2011; 65(9):2421–2434.

740 **Hill WG**, Robertson A. The effect of linkage on limits to artificial selection. Genetics Research. 1966; 8(3):269–294.

741 **Hilliker AJ**, Harauz G, Reaume AG, Gray M, Clark SH, Chovnick A. Meiotic gene conversion tract length distribution
742 within the rosy locus of Drosophila melanogaster. Genetics. 1994; 137(4):1019–1026.

743 **Hinton G**, Deng L, Yu D, Dahl G, rahman Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T,
744 Kingsbury B. Deep Neural Networks for Acoustic Modeling in Speech Recognition. Signal Processing Magazine.
745 2012; .

746 **Hudson RR**. Estimation the recombination parameter of a finite population model without selection. Genetical
747 Research. 1987; 50:245–250.

748 **Hudson RR**. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002
749 Feb; 18(2):337–338.

750 **Hudson RR**, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample
751 of DNA sequences. Genetics. 1985; 111(1):147–164.

752 **Hunter N**, Aguilera A, Rothstein R. Molecular Genetics of Recombination. . 2006; .

753 **Jaenike J**. Sex chromosome meiotic drive. Annual Review of Ecology and Systematics. 2001; 32(1):25–49.

754 **Jozefowicz R**, Zaremba W, Sutskever I. An empirical exploration of recurrent network architectures. In: *Interna-*
755 *tional Conference on Machine Learning*; 2015. p. 2342–2350.

756 **Kelleher J**, Etheridge AM, McVean G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample
757 Sizes. PLOS Computational Biology. 2016 May; 12(5):e1004842. https://doi.org/10.1371/journal.pcbi.1004842,
758 doi: 10.1371/journal.pcbi.1004842.

759 **Kern AD**, Schrider DR. Discoal: flexible coalescent simulations with selection. Bioinformatics. 2016; 32(24):3839–
760 3841.

761 **Kern AD**, Schrider DR. diploS/HIC: An Updated Approach to Classifying Selective Sweeps. G3: Genes, Genomes,
762 Genetics. 2018; 8(6):1959–1970. http://www.g3journal.org/content/8/6/1959, doi: 10.1534/g3.118.200262.

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

763 **Kim Y**, Nielsen R. Linkage disequilibrium as a signature of selective sweeps. Genetics. 2004; 167(3):1513–1524.

764 **Kirkpatrick M**. How and why chromosome inversions evolve. PLoS biology. 2010; 8(9):e1000501.

765 **Kirkpatrick M**, Barton N. Chromosome inversions, local adaptation and speciation. Genetics. 2006; 173(1):419–
766 434.

767 **Kondrashov AS**. Selection against harmful mutations in large sexual and asexual populations. Genetics
768 Research. 1982; 40(3):325–332.

769 **Kong A**, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A,
770 Gylfason A, Kristinsson KT, et al. Fine-scale recombination rate differences between sexes, populations and
771 individuals. Nature. 2010; 467(7319):1099.

772 **Krizhevsky A**, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Net-
773 works. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. *Advances in Neural Informa-*
774 *tion Processing Systems 25* Curran Associates, Inc.; 2012.p. 1097–1105. http://papers.nips.cc/paper/
775 4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

776 **Kulathinal RJ**, Stevison LS, Noor MA. The genomics of speciation in Drosophila: diversity, divergence, and
777 introgression estimated using low-coverage genome sequencing. PLoS genetics. 2009; 5(7):e1000550.

778 **Lack JB**, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, Langley CH, Pool JE. The Drosophila
779 Genome Nexus: A Population Genomic Resource of 623 Drosophila melanogaster Genomes, Including 197
780 from a Single Ancestral Range Population. Genetics. 2015; 199(4):1229–1241. http://www.genetics.org/
781 content/199/4/1229, doi: 10.1534/genetics.115.174664.

782 **Langley CH**, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, Langley SA, Suarez C, Corbett-Detig RB,
783 Kolaczkowski B, et al. Genomic variation in natural populations of Drosophila melanogaster. Genetics. 2012;
784 192(2):533–598.

785 **Lecun Y**, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. In: *Proceedings*
786 *of the IEEE*; 1998. p. 2278–2324.

787 **Li N**, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-
788 nucleotide polymorphism data. Genetics. 2003; 165(4):2213–2233.

789 **Lin K**, Futschik A, Li H. A fast estimate for the population recombination rate based on regression. Genetics.
790 2013; p. genetics–113.

791 **Liu X**, Fu YX. Exploring population size changes using SNP frequency spectra. Nature Genetics. 2015 04; 47:555
792 EP –. https://doi.org/10.1038/ng.3254.

793 **Lowry DB**, Willis JH. A widespread chromosomal inversion polymorphism contributes to a major life-history
794 transition, local adaptation, and reproductive isolation. PLoS biology. 2010; 8(9):e1000500.

795 **Mather K**. Crossing-over. Biological Reviews. 1938; 13(3):252–292.

796 **McVean G**, Awadalla P, Fearnhead P. A coalescent-based method for detecting and estimating recombination
797 from gene sequences. Genetics. 2002; 160(3):1231–1241.

798 **Miller DE**, Cook KR, Arvanitakis AV, Hawley RS. Third Chromosome Balancer Inversions Disrupt Protein-Coding
799 Genes and Influence Distal Recombination Events in Drosophila melanogaster. G3: Genes, Genomes, Genetics.
800 2016; 6(7):1959–1967. https://www.g3journal.org/content/6/7/1959, doi: 10.1534/g3.116.029330.

801 **Myers S**, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots
802 across the human genome. Science. 2005; 310(5746):321–324.

803 **Myers SR**, Griffiths RC. Bounds on the minimum number of recombination events in a sample history. Genetics.
804 2003; 163(1):375–394.

805 **Noor MA**, Grams KL, Bertucci LA, Reiland J. Chromosomal inversions and the reproductive isolation of species.
806 Proceedings of the National Academy of Sciences. 2001; 98(21):12084–12088.

807 **Novitski E**, Braver G. An analysis of crossing over within a heterozygous inversion in Drosophila melanogaster.
808 Genetics. 1954; 39(2):197.

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

O'Reilly PF, Birney E, Balding DJ. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. Genome research. 2008; 18(8):1304–1313.

Parsch J, Meiklejohn CD, Hartl DL. Patterns of DNA sequence variation suggest the recent action of positive selection in the janus-ocnus region of Drosophila simulans. Genetics. 2001; 159(2):647–657.

Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchen P, Emerson JJ, Saelao P, Begun DJ, Langley CH. Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. PLOS Genetics. 2012 12; 8(12):1–24. https://doi.org/10.1371/journal.pgen.1003080, doi: 10.1371/journal.pgen.1003080.

Presgraves DC, Gérard PR, Cherukuri A, Lyttle TW. Large-scale selective sweep among segregation distorter chromosomes in African populations of Drosophila melanogaster. PLoS genetics. 2009; 5(5):e1000463.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS genetics. 2009; 5(6):e1000519.

Przeworski M, Wall JD. Why is there so little intragenic linkage disequilibrium in humans? Genetics Research. 2001; 77(2):143–151.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria; 2018, https://www.R-project.org.

Rieseberg LH. Chromosomal rearrangements and speciation. Trends in ecology & evolution. 2001; 16(7):351–358.

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

Rogers AR. How population growth affects linkage disequilibrium. Genetics. 2014; 197(4):1329–1341.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vision. 2015 Dec; 115(3):211–252. http://dx.doi.org/10.1007/s11263-015-0816-y, doi: 10.1007/s11263-015-0816-y.

Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nature Genetics. 2014 06; 46:919 EP –. https://doi.org/10.1038/ng.3015.

Schrider DR, Ayroles J, Matute DR, Kern AD. Supervised machine learning reveals introgressed loci in the genomes of Drosophila simulans and D. sechellia. PLoS genetics. 2018; 14(4):e1007341.

Schrider DR, Kern AD. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends in Genetics. 2018 Apr; 34(4):301–312. https://doi.org/10.1016/j.tig.2017.12.005, doi: 10.1016/j.tig.2017.12.005.

Schrider DR, Mendes FK, Hahn MW, Kern AD. Soft shoulders ahead: spurious signatures of soft and partial selective sweeps result from linked hard sweeps. Genetics. 2015; 200(1):267–284.

Schumer M, Xu C, Powell DL, Durvasula A, Skov L, Holland C, Blazier JC, Sankararaman S, Andolfatto P, Rosenthal GG, Przeworski M. Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science. 2018; 360(6389):656–660. https://science.sciencemag.org/content/360/6389/656, doi: 10.1126/science.aar3684.

Singh ND, Stone EA, Aquadro CF, Clark AG. Fine-scale heterogeneity in crossover rate in the garnet-scalloped region of the Drosophila melanogaster X chromosome. Genetics. 2013; 194(2):375–387.

Slatkin M. Linkage disequilibrium in growing and stable populations. Genetics. 1994; 137(1):331–336.

Smith KN, Nicolas A. Recombination at work for meiosis. Current opinion in genetics & development. 1998; 8(2):200–211.

Sturtevant A. A case of rearrangement of genes in Drosophila. Proceedings of the National Academy of Sciences of the United States of America. 1921; 7(8):235.

Sutskever I, Vinyals O, Le QV. Sequence to Sequence Learning with Neural Networks. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* NIPS'14, Cambridge, MA, USA: MIT Press; 2014. p. 3104–3112. http://dl.acm.org/citation.cfm?id=2969033.2969173.

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-4.0 International license.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed SE, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*; 2015. p. 1–9. https://doi.org/10.1109/CVPR.2015.7298594, doi: 10.1109/CVPR.2015.7298594.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, et al. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science. 2012; 337(6090):64–69. https://science.sciencemag.org/content/337/6090/64, doi: 10.1126/science.1219240.

Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nature Genetics. 2016 12; 49:303 EP –. https://doi.org/10.1038/ng.3748.

Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and Composing Robust Features with Denoising Autoencoders. In: *Proceedings of the 25th International Conference on Machine Learning* ICML '08, New York, NY, USA: ACM; 2008. p. 1096–1103. http://doi.acm.org/10.1145/1390156.1390294, doi: 10.1145/1390156.1390294.

Wakeley J. Using the variance of pairwise differences to estimate the recombination rate. Genetics Research. 1997; 69(1):45–48.

Wall JD. A comparison of estimators of the population recombination rate. Molecular Biology and Evolution. 2000; 17(1):156–163.

Wang RJ, Gray MM, Parmenter MD, Broman KW, Payseur BA. Recombination rate variation in mice from an isolated island. Molecular ecology. 2017; 26(2):457–470.

Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, McVean GA, Gabriel SB, Reich D, Donnelly P, et al. Comparison of fine-scale recombination rates in humans and chimpanzees. Science. 2005; 308(5718):107–111.

Wiuf C. On the minimum number of topologies explaining a sample of DNA sequences. Theoretical population biology. 2002; 62(4):357–363.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, Zheng H, Liu T, He W, Li K, Luo R, Nie X, Wu H, Zhao M, Cao H, Zou J, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. Science (New York, NY). 2010 07; 329(5987):75–78. https://www.ncbi.nlm.nih.gov/pubmed/20595611, doi: 10.1126/science.1190371.

Zelkowski M, Olson M, Wang M, P Pawlowski W. Diversity and Determinants of Meiotic Recombination Landscapes. Trends in Genetics. 2019 04; doi: 10.1016/j.tig.2019.02.002.

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-4.0 International license.

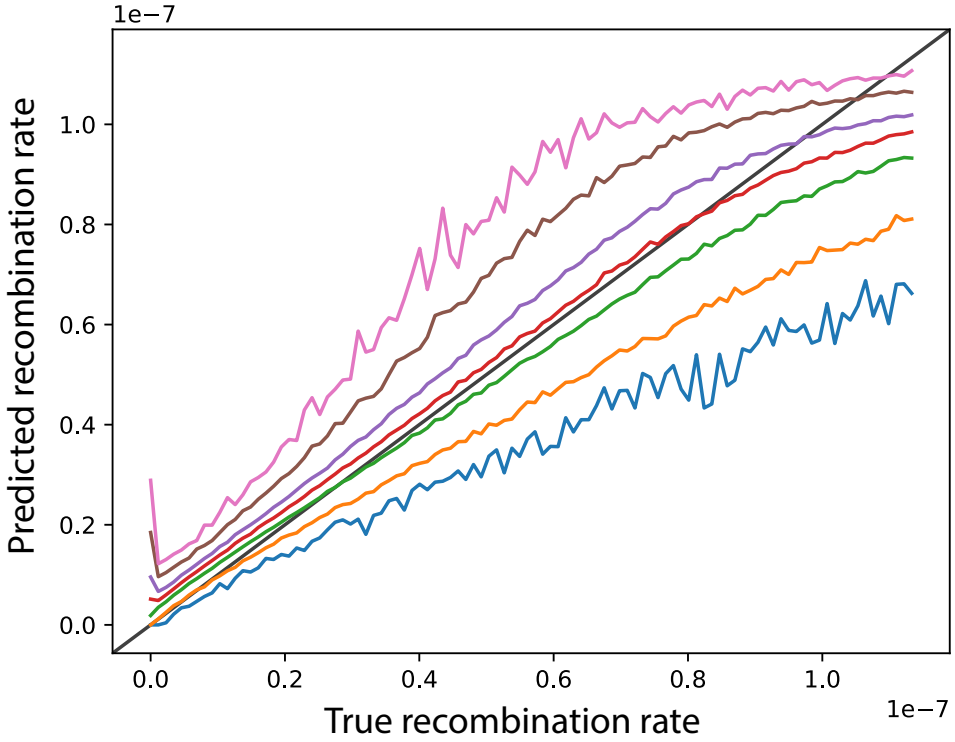

Figure 1–Figure supplement 1. Parametric bootstraping results as implemented by ReLERNN. Lines represent the minimum (blue), lower 5% (orange), lower 25% (green), median (red), upper 25% (purple), upper 95% (brown), and maximum (pink) bounds for each of 1000 replicate simulations and predictions (y-axis) across 100 recombination rate bins (x-axis)

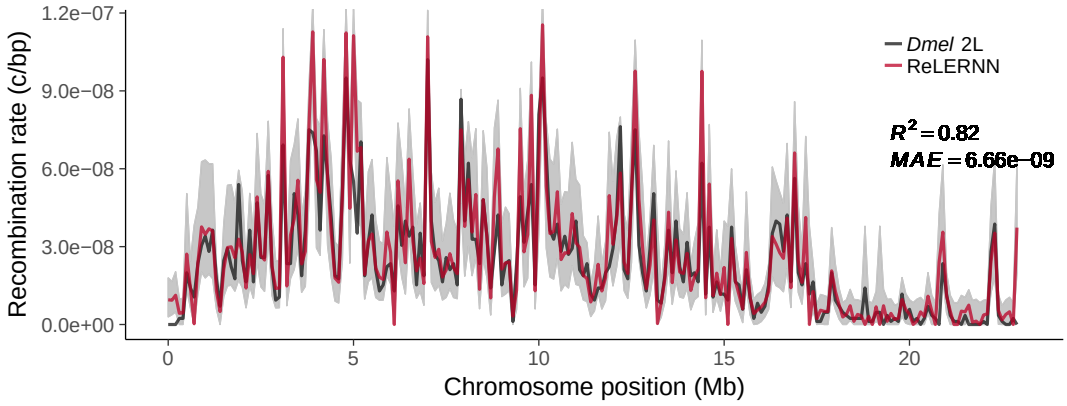

Figure 2–Figure supplement 1. Recombination rate predictions for a simulated *Drosophila* chromosome (black line) using ReLERNN (red line). The recombination landscape was simulated for $n = 4$ chromosomes under mutation-drift equilibrium using msprime (***Kelleher et al., 2016***), with per-base crossover rates derived from *D. melanogaster* chromosome 2L (***Comeron et al., 2012***). Gray ribbons represent 95% confidence intervals. R^2 is reported for the general linear model of predicted rates on true rates and mean absolute error was calculated across all 100 kb windows.

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-4.0 International license.

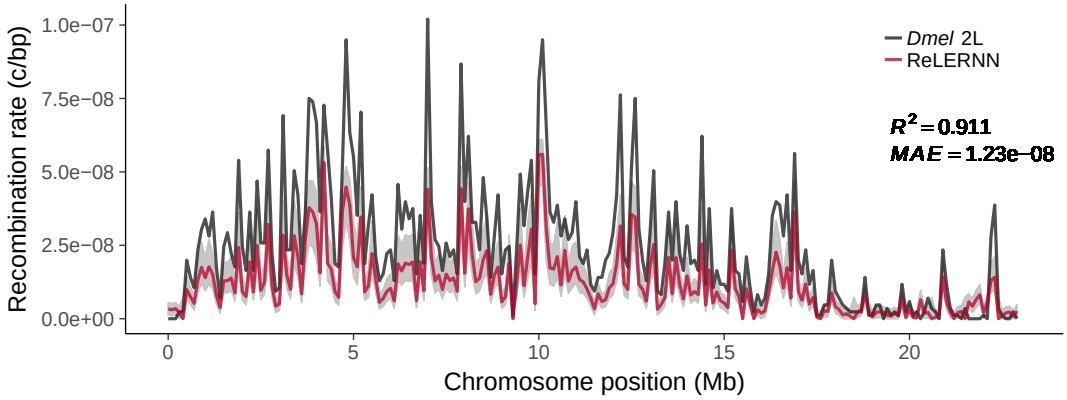

Figure 2–Figure supplement 2. Recombination rate predictions for a simulated *Drosophila* chromosome (black line) using ReLERNN (red line). The recombination landscape was simulated for $n = 20$ chromosomes under mutation-drift equilibrium using msprime (***Kelleher et al., 2016***), with per-base crossover rates derived from *D. melanogaster* chromosome 2L (***Comeron et al., 2012***). Here the per-base mutation rate was assumed to be 50% less than the rate used for simulation. Gray ribbons represent 95% confidence intervals. R^2 is reported for the general linear model of predicted rates on true rates and mean absolute error was calculated across all 100 kb windows.

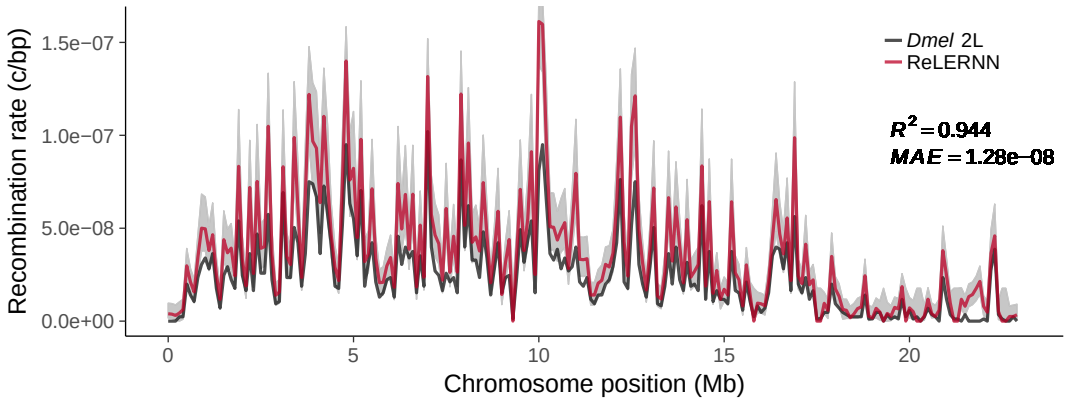

Figure 2–Figure supplement 3. Recombination rate predictions for a simulated *Drosophila* chromosome (black line) using ReLERNN (red line). The recombination landscape was simulated for $n = 20$ chromosomes under mutation-drift equilibrium using msprime (***Kelleher et al., 2016***), with per-base crossover rates derived from *D. melanogaster* chromosome 2L (***Comeron et al., 2012***). Here the per-base mutation rate was assumed to be 50% greater than the rate used for simulation. Gray ribbons represent 95% confidence intervals. R^2 is reported for the general linear model of predicted rates on true rates and mean absolute error was calculated across all 100 kb windows.

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

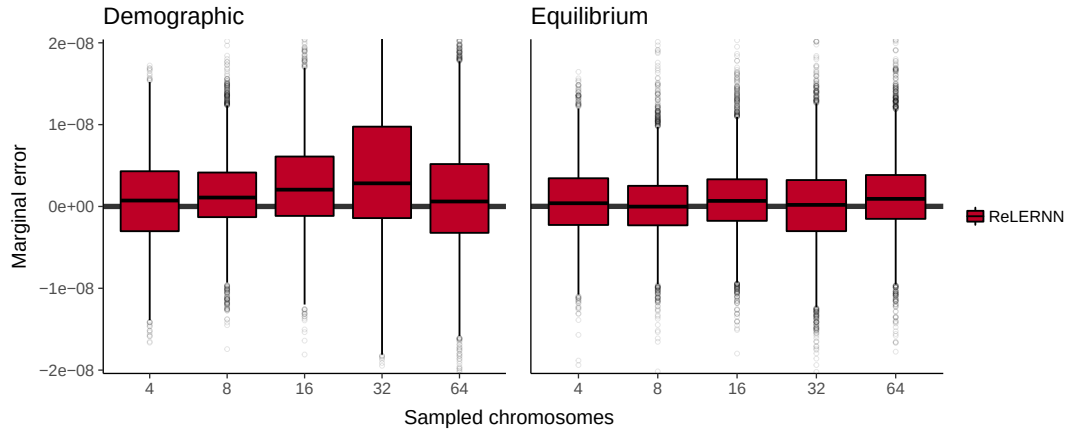

Figure 4–Figure supplement 1. Distribution of marginal errors attributed to model misspecification across 5000 simulated chromosomes. Predictions were made by training on equilibrium simulations and testing on sequences simulated under a demographic model (left) or training on demographic simulations and testing on sequences simulated under equilibrium (right). Here, marginal errors are represented as $\epsilon_m - \epsilon_c$, where ϵ_m and ϵ_c are equal to $|r_{predicted} - r_{true}|$ when the model is misspecified and correctly specified, respectively. Sampled chromosomes indicate the number of independent sequences that were sampled from each msprime (***Kelleher et al., 2016***) coalescent simulation.

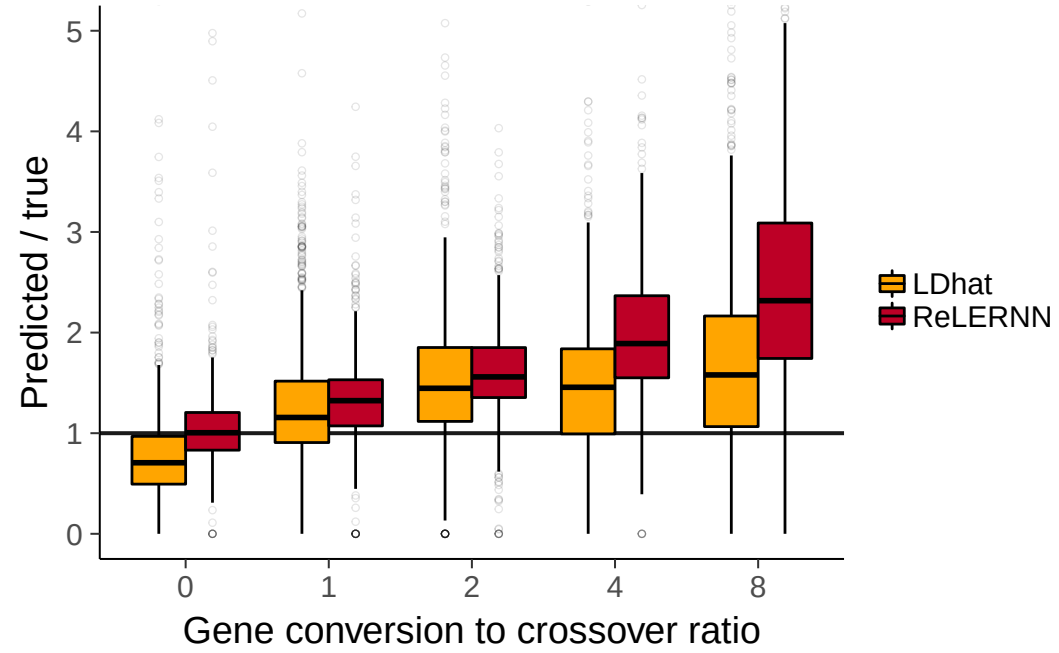

Figure 4–Figure supplement 2. Distribution of predicted rates of recombination over true rates for 5000 examples simulated with gene conversion and $n = 8$. The ratio of gene conversion to crossovers was drawn from $U(0, c)$, with $c \in \{0, 1, 2, 4, 8\}$. Gene conversion tract lengths were fixed at 352 bp, and all simulations were completed in ms (***Hudson, 2002***).

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-4.0 International license.

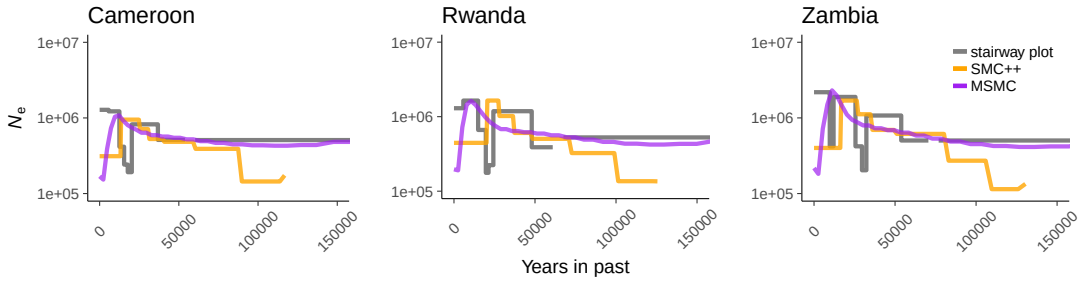

Figure 5–Figure supplement 1. Historical population size estimates were inferred for Cameroon, Rwanda, and Zambia using three separate methods, all of which disagree with one another. Inferences are based on 10 samples for both stairwayplot (grey line) and SMC++ (orange line), and 2 samples for MSMC (purple line).

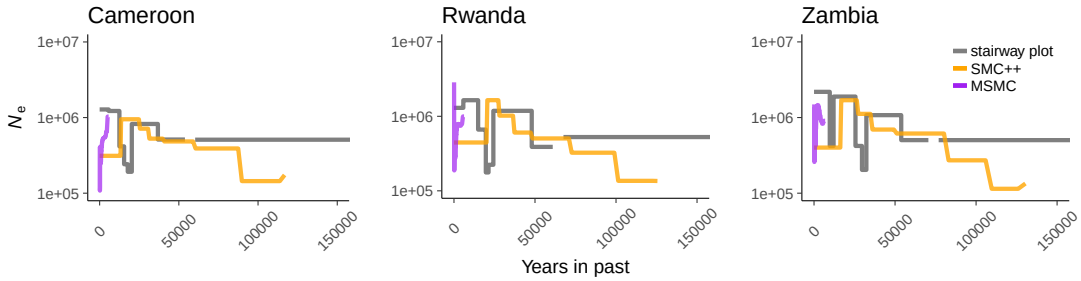

Figure 5–Figure supplement 2. Historical population size estimates were inferred for Cameroon, Rwanda, and Zambia using three separate methods, all of which disagree with one another. Here, inferences are based on 10 samples for both stairwayplot (grey line) and SMC++ (orange line), and 10 samples for MSMC (purple line).

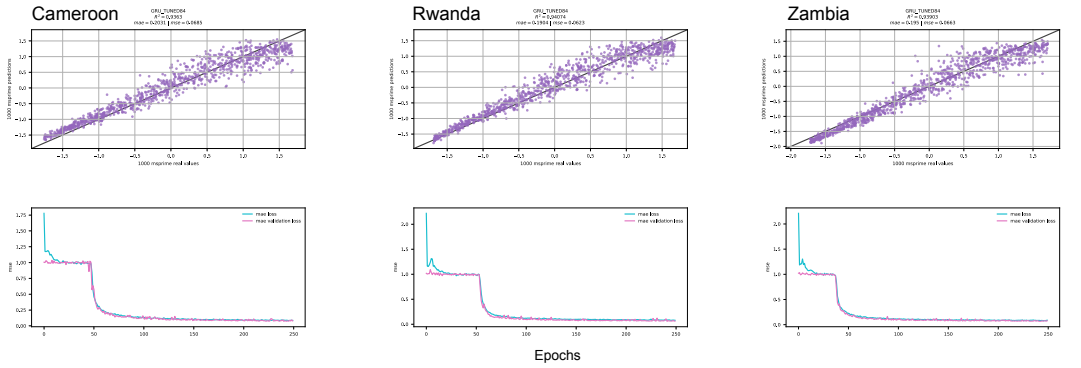

Figure 5–Figure supplement 3. ReLERNN test results for Cameroon, Rwanda, and Zambia when trained under assumptions of mutation-drift equilibrium. Scatter plots (top) show raw (unnormalized) predictions for per-base recombination rates for 1000 test examples. Mean absolute error and mean squared error are calculated for each population. Line graphs (bottom) show the decrease in the mean absolute error over time (epochs) for both the training set (blue lines) and the validation set (purple lines).

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

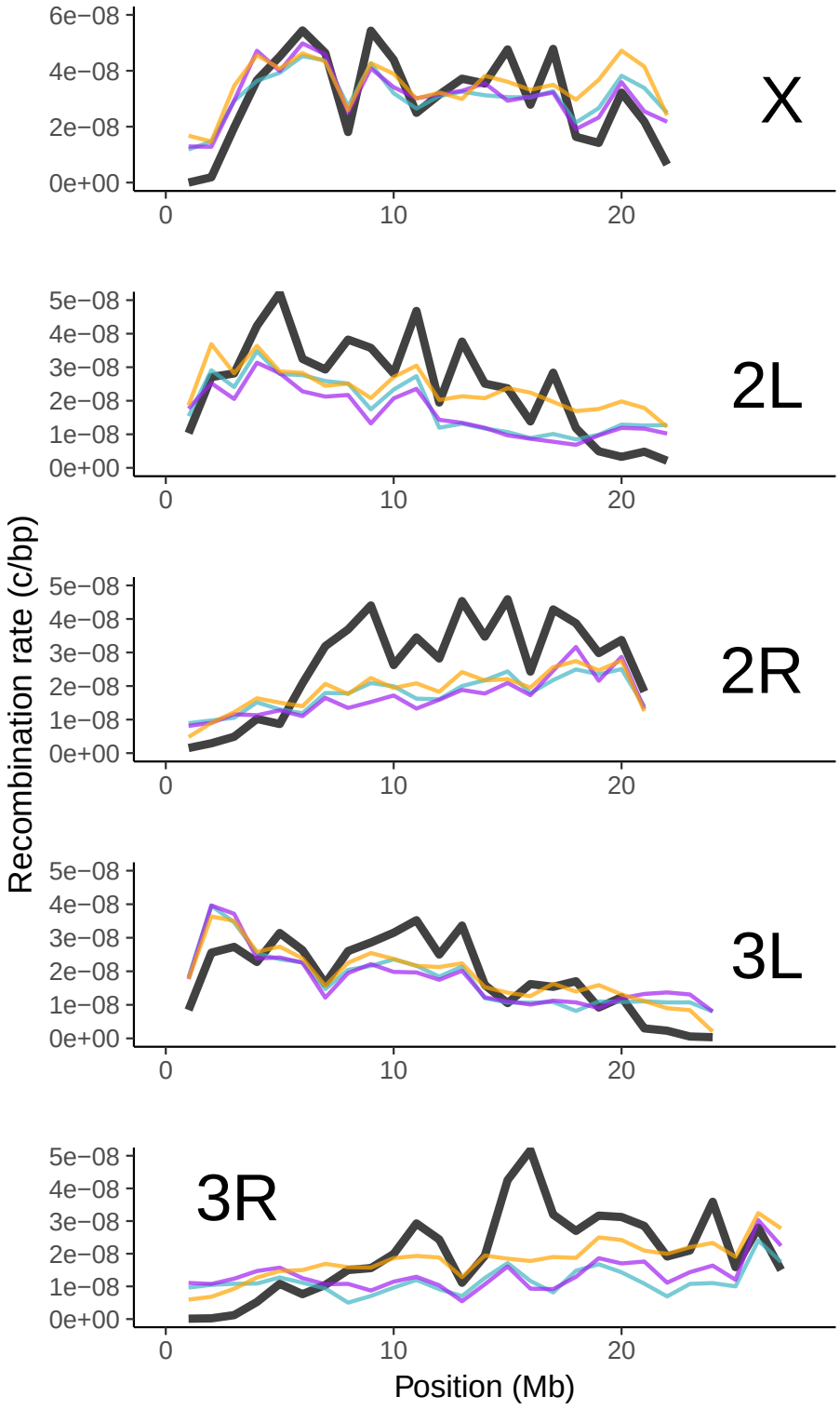

Figure 5–Figure supplement 4. Genome-wide recombination landscapes for *D. melanogaster* populations from Cameroon (teal lines), Rwanda (purple lines), and Zambia (orange lines). Rates are compared to those experimentally derived by ***Comeron et al.*** (***2012***) (black lines). All rates have been scales to 1 Mb windows by using a weighted average (see Materials and Methods).

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

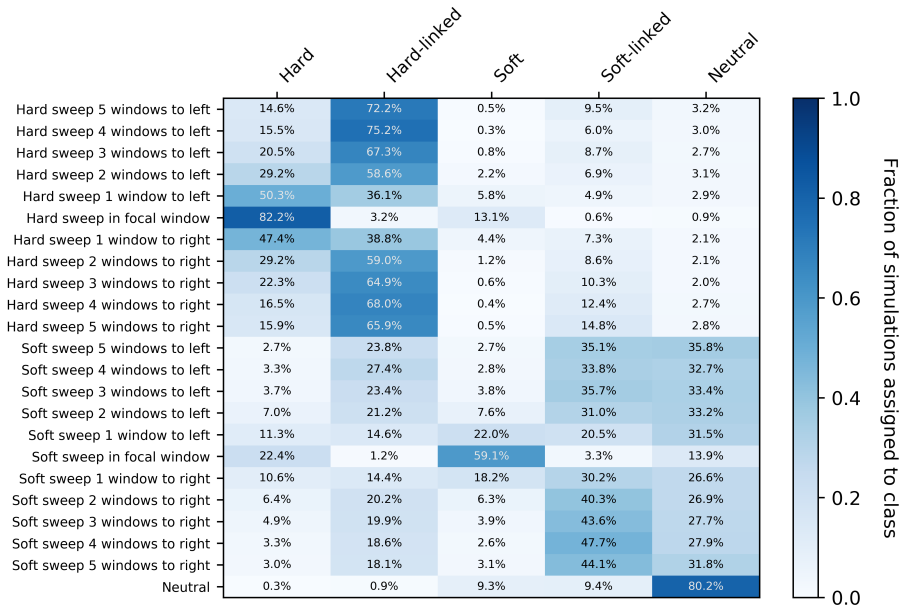

Figure 5–Figure supplement 5. Confusion matrix showing the fraction of test simulation windows assigned to each of five prediction categories by diploS/HIC (***Kern and Schrider, 2018***): hard, hard-linked, soft, soft-linked, and neutral. The y-axis shows the location of the window being classified relative to the selected window.

bioRxiv preprint doi: https://doi.org/10.1101/662247; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

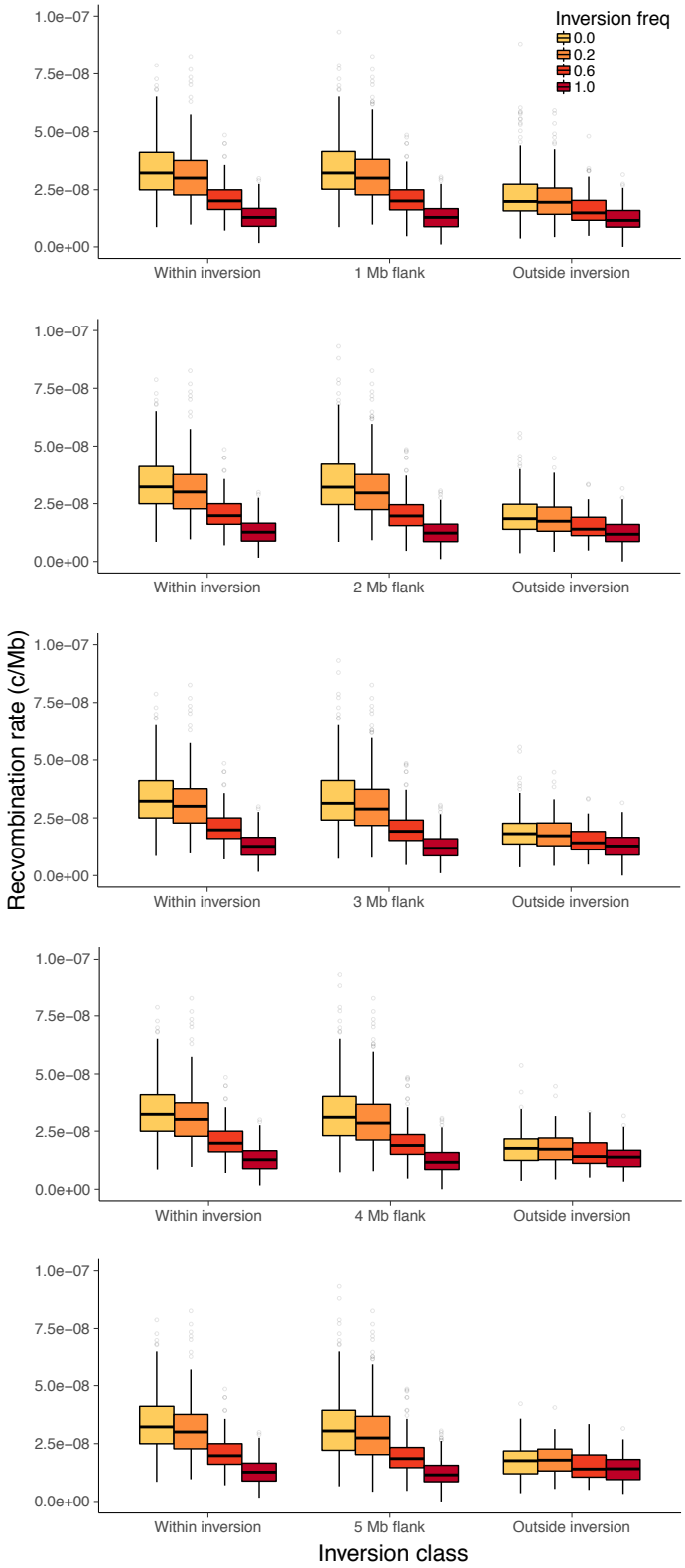

Figure 6–Figure supplement 1. Recombination rate estimates using flanking window sizes from 1-5 Mb. Rates are shown for genomic windows within the inversion, within regions flanking the inversion, and for regions outside both the inversion and flanking regions. All estimates are from chromosome 2L with *In(2L)t* sampled at different inversion frequencies