# The Theory and Applications of Measuring Broad-Range and Chromosome-Wide Recombination Rate from Allele Frequency Decay around a Selected Locus

Kevin H.-C. Wei,* Aditya Mantha, and Doris Bachtrog

Department of Integrative Biology, University of California Berkeley, Berkeley, CA

*Corresponding author: E-mail: weikevin@berkeley.edu.

**Associate editor:** Nadia Singh

All generated short reads can be found on the Sequenced Reads Archive under BioProject: PRJNA564642.

## Abstract

**Recombination is the exchange of genetic material between homologous chromosomes via physical crossovers. High-throughput sequencing approaches detect crossovers genome wide to produce recombination rate maps but are difficult to scale as they require large numbers of recombinants individually sequenced. We present a simple and scalable pooled-sequencing approach to experimentally infer near chromosome-wide recombination rates by taking advantage of non-Mendelian allele frequency generated from a fitness differential at a locus under selection. As more crossovers decouple the selected locus from distal loci, the distorted allele frequency attenuates distally toward Mendelian and can be used to estimate the genetic distance. Here, we use marker selection to generate distorted allele frequency and theoretically derive the mathematical relationships between allele frequency attenuation, genetic distance, and recombination rate in marker-selected pools. We implemented nonlinear curve-fitting methods that robustly estimate the allele frequency decay from batch sequencing of pooled individuals and derive chromosome-wide genetic distance and recombination rates. Empirically, we show that marker-selected pools closely recapitulate genetic distances inferred from scoring recombinants. Using this method, we generated novel recombination rate maps of three wild-derived strains of *Drosophila melanogaster*, which strongly correlate with previous measurements. Moreover, we show that this approach can be extended to estimate chromosome-wide crossover interference with reciprocal marker selection and discuss how it can be applied in the absence of visible markers. Altogether, we find that our method is a simple and cost-effective approach to generate chromosome-wide recombination rate maps requiring only one or two libraries.**

***Key words:*** **recombination, crossovers, pooled sequencing, allele frequency.**

## Introduction

T.H. Morgan first envisioned the exchange of genetic material between homologous chromosomes through physical crossovers (Morgan 1911b). After backcrossing F1 heterozygotes, he recovered novel allelic combinations, or recombinants, of markers on the same chromosome that are absent in parental lines (Morgan 1911a). At face value, this appeared to violate Mendel's law of segregation and the chromosome theory of inheritance (Blixt 1975). A.H. Sturtevant, then a prodigious undergraduate in Morgan's laboratory, confirmed his mentor's theory by crossing mutant *Drosophila* strains with different visible X-linked markers. Noticing that different pairs of markers produced recombinants at different frequencies, Sturtevant realized that the positions of these markers can be ordered on a linear map, with the frequency of recombinants as the genetic distance between any two markers (Sturtevant 1913). This was the birth of the very first genetic map (Brush 2002).

Measured as the number of recombinants divided by the total number of offsprings, the frequency of recombinants (henceforth recombinant fraction) does not reflect the true probability of crossovers between two loci (Castle 1919). This is because only odd numbers of crossovers are observable since even numbers of crossovers produce recombinants that maintain the allelic combinations of the parents (Sturtevant 1913; Sturtevant et al. 1919). With increasing genetic distance, the probability of multiple crossovers also increases causing recombinant fraction to deviate from true genetic distance. To account for multiple crossovers, mapping functions are applied to convert recombinant fractions ($D$) to true genetic distances ($d$). Although many mapping functions have been developed to account for different rates of crossover interference (Felsenstein 1979; McPeek and Speed 1995; Tan and Fornage 2008), the two most popular are Haldane's (Haldane 1919) and Kosambi's mapping functions (Kosambi 1943). Haldane's mapping function (**H**) assumes no crossover interference and tends to overestimate genetic distance

(Tan and Fornage 2008). Kosambi's mapping function (**K**) assumes that crossover interference is inversely proportional to the genetic distance between any two loci and produces estimates that are more consistent with crossing experiments (Huehn 2011). After these functions, the genetic distance is expected to be additive such that $\mathbf{H}(D_{AC}) = \mathbf{H}(D_{AB}) + \mathbf{H}(D_{BC})$ given three loci A, B, and C, in that order.

In the century after Morgan and Sturtevant conceptualized and devised the recombinant backcross scheme, recombination has been recognized as one of the most fundamental and universal biological processes across sexually reproducing eukaryotes with vast implications in many biological fields. For example, crossover events are crucial for the fidelity of chromosome segregation (see reviews, Page and Hawley 2004; Hunter 2015; Hughes et al. 2018) and recombination rate is intimately linked with the efficacy of natural selection and genome evolution (see reviews, Cutter and Payseur 2013; Martin and Jiggins 2017; Stephan 2019). Despite rapid technological and methodological advances, methods to measure recombinant fraction, genetic distance, and recombination rate remain grounded in the approach devised by Sturtevant—tallying the number of recombinants between loci that are either phenotypically or molecularly marked. The advent of whole-genome sequencing has permitted the generation of high-resolution crossover maps, by identifying regions of the genome where parental haplotypes change in recombinant individuals (Kulathinal et al. 2008; Rockman and Kruglyak 2009; Dumont et al. 2011; Comeron et al. 2012; Miller et al. 2012). More recently, single-cell sequencing technologies have further extended such high-throughput crossover detection directly from sperm cells (Hinch et al. 2019). Although providing impressive resolution often at the level of individual bases, these approaches are difficult to scale, as each recombinant individual/cell requires a separate library preparation and/or barcode. Since each chromosome, on average, has only one to two crossovers, hundreds to thousands of individuals, and therefore libraries/barcodes, need to be sequenced for a comprehensive map.

In addition to cross-based experimental measurements, recombination rate is also frequently estimated at the population level (Langley et al. 2000; McVean et al. 2004; Wang and Rannala 2008; Chan et al. 2012). Such population estimates, in short, are inferred from the breakdown of linkage patterns from historical recombination events. Recent developments have further allowed highly cost-effective and accurate population level estimates from single diploid genomes (Barroso et al. 2019) and pooled sequencing of unrelated individuals (Adrion et al. 2020). In conjunction with experimental approaches, these methods have been instrumental in revealing the importance of recombination on population dynamics and genome evolution.

However, the lack of a scalable method to experimentally estimate recombination rate at genome scale is particularly problematic since recombination rate is a highly labile phenotype (Ritz et al. 2017). It is sensitive to environmental stresses like temperature (Stern 1926), nutrition (Neel 1941), and infection (Singh et al. 2015), and changes with age (Redfield 1966). Furthermore, recombination rate differs drastically not only between closely related species (Jensen-Seaman et al. 2004; Kulathinal et al. 2008; Smukowski and Noor 2011; Brand et al. 2018) but also between individuals of the same species (Nachman 2002; Dumont et al. 2009, 2011; Stevison and Noor 2010; Comeron et al. 2012; Kaur and Rockman 2014; Hunter et al. 2016). Therefore, a scalable and cost-efficient approach is necessary to fully capture the extent of the variability of recombination rate and enrich our understanding of this fundamental yet volatile molecular process.

Instead of scoring recombinants or identifying crossover breakpoints in individuals from a typical backcross, we demonstrate here that recombination rate can be estimated from allele frequency (AF) attenuation around loci that cause distorted AF. When the homologous alleles across all loci have equal fitness, the AF of the progeny pool is expected to be Mendelian across the chromosome, regardless of the number of recombinants produced by crossovers in the F1 parents (fig. 1*A*). However, when the alleles at a given locus do not have the same fitness/viability, that is, a fitness differential, the AF at the locus will deviate from the Mendelian ratio. Importantly, the deviating AF is expected to show a signature attenuation pattern whereby it peaks at the distorting locus and attenuates distally toward Mendelian due to increasing number of crossover events between the locus and gradually more distal loci. Therefore, the AF and its rate of change should be related to the genetic distance and recombination rate, respectively. Since chromosome-wide AF can be determined using pooled whole-genome sequencing of the backcross progenies (Kofler et al. 2011; Wei et al. 2017; Tilk et al. 2019), the recombination rate map can be generated with as little as one library preparation. This is, at a minimum, two orders of magnitude fewer than current approaches. Fitness differential can be easily achieved by using marker selection, therefore allowing recombination rate to be estimated from AFs in marker-selected pools (fig. 1*B*). Previously, AF has been utilized for fine-scale crossover rate estimates in pools where recombinants were specifically selected by a double marker selection scheme (Singh et al. 2013) but had analytical shortfalls that resulted in problematic estimates (Gilliland 2015) (also see Discussion). Our approach here requires only one marker and does not require identification and scoring of visible recombinants.

We theoretically, computationally, and empirically explore the series of steps to infer recombination rates from AF attenuation in marker-selected pools. First, we formally demonstrate the mathematical relationship between recombination rate and AF using the *Drosophila* X chromosome with the recessive white eye marker (w-) as an example. Then, we generalize the relationships allowing for applications in the presence of any locus with a fitness differential and demonstrate how additive fitness impacts from offsite loci modulate the AF. Using simulations, we show how inherent noise in the sequencing platform can be addressed statistically for robust AF estimates. For empirical validation, we show that genetic distances estimated from the AF changes closely recapitulate the distances estimated from scoring recombinant individuals.

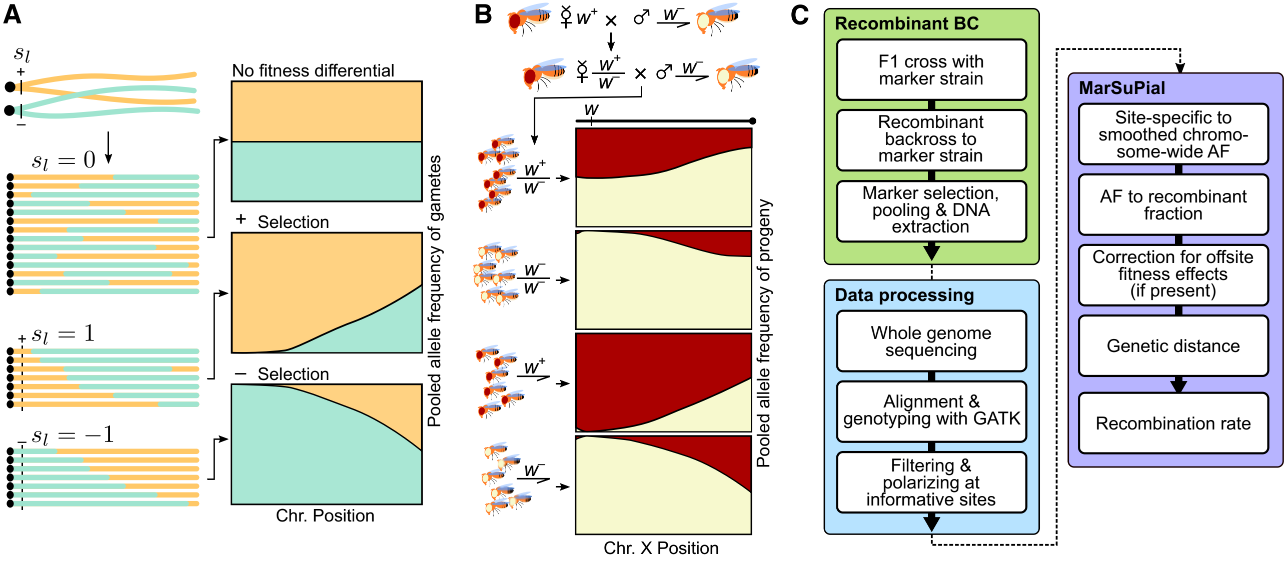
A
$s_l$
$s_l = 0$
$s_l = 1$
$s_l = -1$
No fitness differential
+ Selection
− Selection
Pooled allele frequency of gametes
Chr. Position
B
♀ $w^+$ × ♂ $w^-$
♀ $\frac{w^+}{w^-}$ × ♂ $w^-$
w
$\frac{w^+}{w^-}$
$\frac{w^-}{w^-}$
$w^+$
$w^-$
Pooled allele frequency of progeny
Chr. X Position
C
Recombinant BC
F1 cross with marker strain
Recombinant backross to marker strain
Marker selection, pooling & DNA extraction
Data processing
Whole genome sequencing
Alignment & genotyping with GATK
Filtering & polarizing at informative sites
MarSuPial
Site-specific to smoothed chromosome-wide AF
AF to recombinant fraction
Correction for offsite fitness effects (if present)
Genetic distance
Recombination rate

FIG. 1. Estimating recombination rate around locus with fitness differential. (*A*) Conceptual schematic of AF attenuation around a locus with fitness differential. On the left are cartoon schematics of crossover between homologous chromosomes and their recombinant products. After strong sources of selection (with markers) at locus $s_l$ (dotted line), the recombinant chromosomes are ordered by the length of the haplotypes for illustrative purposes. The AFs of the recombinant pools are depicted in the cartoons to the right. (*B*) Application of this approach using the X-linked eye color marker white (*w*). The recombinant backcross scheme is on the top and the resulting four sexed and genotyped BC1 pools are displayed on the bottom. The AFs of these pools are displayed to the right, where the red and white areas represent the frequency of the $w^+$ and $w^-$ alleles, respectively. (*C*) Workflow of the method, from crosses to allele counting to recombination rate estimation using the Software package MarSuPial.

Using this approach, we generated the third chromosome recombination map of three wild-derived strains of *Drosophila melanogaster*, producing rates that are highly correlated with previous estimates. Lastly, we extend this approach to estimate chromosome-wide crossover interference. Although this marker selection and sequencing scheme does not produce basepair resolution crossover maps since it does not infer breakpoints, it can generate a near chromosome-wide genetic and recombination rate maps in as little as one or two library preparations.

## Results

### Recombinant Fraction Is Directly Proportional to AF in Marker-Selected Pools

In a typical recombinant backcross, F1 heterozygotes are backcrossed to one of the inbred parents generating BC1 progenies that are scored for recombinants (fig. 1A). Although gametes of the F1s are ideal for detecting crossovers as they are direct products of meiosis (Hinch et al. 2019), there are currently very few viable and reliable means of gamete selection (Umehara et al. 2019). Marker selection in progenies entails that the half of the genome derived from the inbred backcross parent (paternal genome in flies) is uninformative despite contributing to the AF. Since the X chromosome in sons are maternally inherited, male and female pools will have different AF signatures with the male pool reflecting the gametic AF (fig. 1B). For simplicity, we will first focus on the male marker-selected pools as their X chromosomes are free of paternal contributions. For illustrative purposes, we will use the X-linked *white* gene (*w*) in *Drosophila* as the selected locus (fig. 1B); selection is based on the recessive phenotype of white eyes ($w^-$) versus red eyes ($w^+$). Between the selected locus *w* and any position *i*, the recombinant fraction between the two loci is $D_i$; the frequency that the $w^-$ and $w^+$ chromosomes had zero or even numbers of crossovers between *w* and *i* ($p_{\text{even}i}$ and $q_{\text{even}i}$, respectively) is $p_{\text{even}i} = (1 - D_i)/2$ and $q_{\text{even}i} = (1 - D_i)/2$. The frequency of odd numbers of crossovers is then $p_{\text{odd}i} = D_i/2$ and $q_{\text{odd}i} = D_i/2$. In the absence of selection, the AF of the allele on the $w^-$ chromosome ($q$) will therefore be $q_i = q_{\text{even}i} + p_{\text{odd}i} = (1 - D_i)/2 + D_i/2 = 0.5$. When $w^-$ is selected, $p_{\text{odd}i} = 0$, therefore,

$$q_i = 1 - D_i \tag{1}$$

and when $w^+$ is selected,

$$q_i = D_i. \tag{2}$$

Thus, the AF is directly proportional to the recombinant fraction in marker-selected pools.

To illustrate this relationship between recombination and AF change in the male selected pools, we used the available X chromosome recombination rate from the Recombination Rate Calculator (Fiston-Lavier et al. 2010) which models the chromosome-wide rate by a quadratic function. We converted the recombination rate to genetic distance from the white locus, followed by transformation into the recombinant fraction using either the Haldane or Kosambi mapping functions (fig. 2A). Depending on whether $w^-$ and $w^+$ males are selected, AF peaks ($q = 1$) or troughs ($q = 0$) at the selected white locus, respectively,

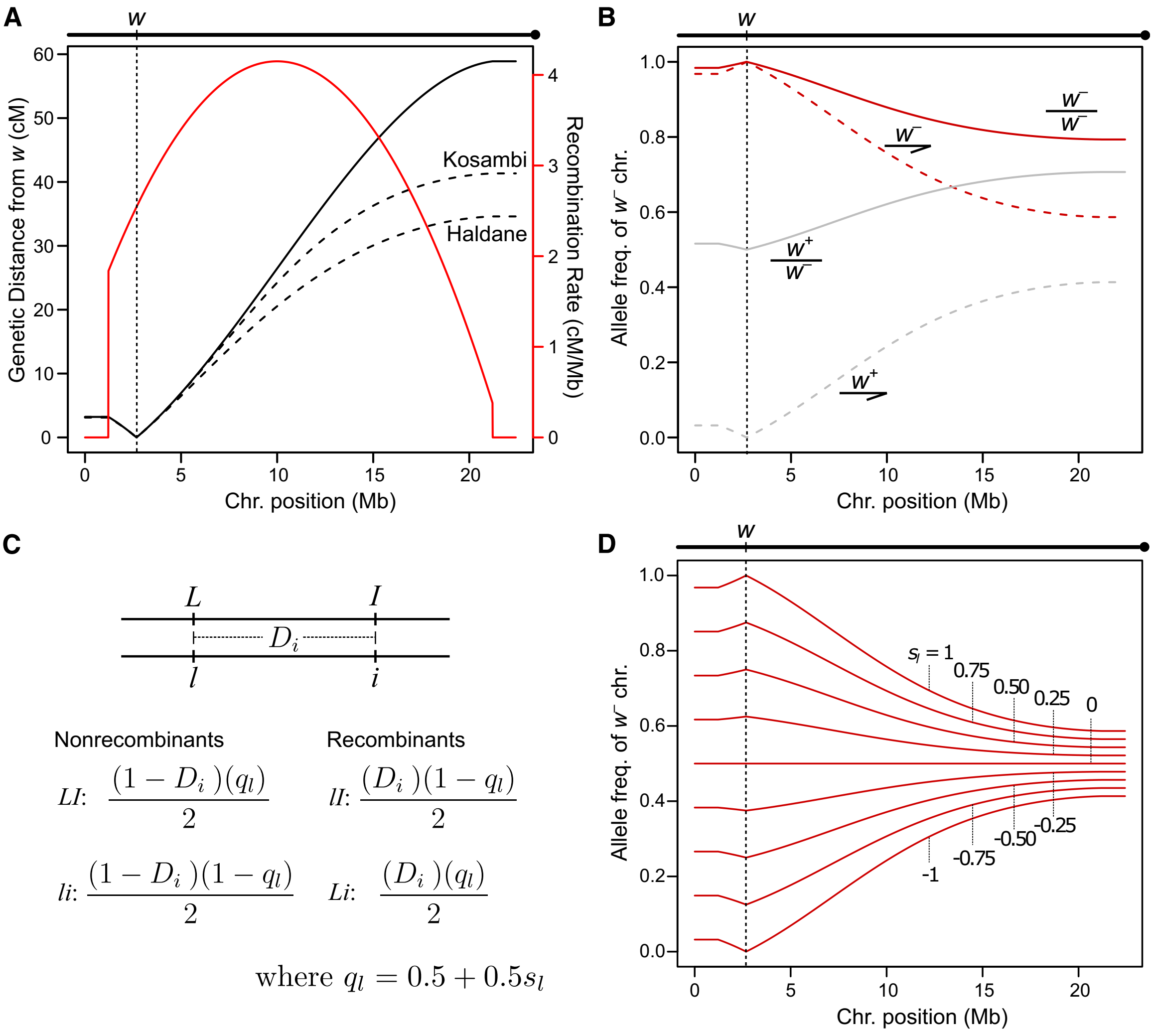

**Fig. 2.** Theoretical relationship between recombination rate, genetic distance, recombinant fraction, selection differential, and AF attenuation. (*A*) Based on the backcross scheme in figure 1*B*, transformation between the recombination rate of the X chromosome (red line), the genetic distance from the white locus (solid black line), and the recombinant fractions after applying the Kosambi and Haldane mapping functions. (*B*) AF of different pools across the X chromosome given the Kosambi transformed recombinant fraction from A. The $w^-$ and $w^+$ pools are in gray and red lines, respectively; the female and male pools are in solid and dotted lines, respectively. (*C*) Crossover schematic between homologous chromosomes carrying linked alleles *L* and *I* their respective homologs and *l* and *i*; the two loci have a distance of $D_i$. Below the schematic, the frequencies of all four possible allelic combinations are depicted, with the nonrecombinants on the left and the recombinants on the right. (*D*) AF attenuation based on different fitness differential ($s_l$).

and attenuates distally, approaching the expected Mendelian ratio of 0.5 (fig. 2*B*).

In order to select for the recessive $w^-$ marker in females, they must be sired by $w^-$ fathers (fig. 1*B*). In the homozygous female ($w^-/w^-$) marker-selected pool, the $w^-$ chromosome AF with paternal contributions will therefore be

$$q_i = \frac{1 - D_i + 1}{2}. \qquad (3)$$

The AF peaks ($q = 1$) at the selected white locus and attenuates toward the expected Mendelian ratio of $q = 0.75$ (fig. 2*B*). If, instead, $w^+$ heterozygous females ($w^+/w^-$) are selected and pooled, the AF troughs at $q = 0.5$ (fig. 2*B*). Autosomal marker selection will have the same AF attenuation pattern.

## Selection/Fitness Differential Creates AF Deviation and Decay

The relationship between *D* and *q* can be further generalized to any locus (*l*) where the two alleles have differential fitness of *s* ranging between −1 and 1, such that

$$q_l = 0.5 + 0.5s_l. \qquad (4)$$

Then, in the absence of paternal contribution, the AF at position *i* is captured by the addition of the proportion of alleles that did not recombine between *l* and *i* and proportion in which recombination occurred between *l* and *i* (fig. 2*C*):

$$q_i = (1 - D_i)(q_l) + (D_i)(1 - q_l). \qquad (5)$$

When $s_l = -1$ and $s_l = 1$, $q_l$ equals 0 and 1, respectively, which recapitulates marker selection as described in equations (1) and (2). When $s_l = 0$, that is, no difference in fitness between alleles, the AF is Mendelian ($q = 0.5$). With other values of $s$, the AF still peaks and decays around $l$ (fig. 2D), thus allowing for estimation of $D$. Therefore, even in the absence of a visible and selectable marker, as long as a locus exists where alleles have differential fitness causing non-Mendelian AF, recombination rate can be estimated around it. Such a locus may be a naturally segregating deleterious or lethal alleles (McCune et al. 2002), a meiotic driver (Fishman and Willis 2005; Wei et al. 2017), or even a partially penetrant marker, allowing for wide applicability of this approach beyond model organisms.

### Accounting for Offsite Viability Effects That Modulate AF Decay

Although the AF decay pattern is a product of crossovers between the selected and distal loci, it is also sensitive to loci with alleles that differentially affect viability, body size, and fitness (which we will collectively refer to as viability effects), as such alleles can cause non-Mendelian contributions of individuals and DNA in the final pool. To evaluate how such loci will affect AF estimates, we further expand on the mathematical relationship between AF and recombinant fraction (eqs. 1–3) by a secondary locus at site $o$ that also modulates AF in addition to the selected locus at site $l$. For simplicity, we are removing the paternal contribution. The two loci have respective viability effect of $s_o$ and $s_l$ that individually (absent of other contributing loci) cause AF ($\widehat{q}$) of $\widehat{q}_l = 0.5 + 0.5s_l$ and $\widehat{q}_o = 0.5 + 0.5s_o$. Given that the distance between $l$ and $o$ is $D_{lo}$, the AF at $o$ is then

$$q_o = \frac{(1-D_{lo})\widehat{q}_l\widehat{q}_o + D_{lo}(1-\widehat{q}_l)(\widehat{q}_o)}{(1-D_{lo})\widehat{q}_l\widehat{q}_o + (1-D_{lo})(1-\widehat{q}_l)(1-\widehat{q}_o) + D_{lo}(1-\widehat{q}_l)(\widehat{q}_o) + D_{lo}(\widehat{q}_l)(1-\widehat{q}_o)}.$$

With the selection process acting at $l$, $s_l = 1$ and $\widehat{q}_l = 1$, so the equation reduces to

$$q_o = \frac{(1-D_{lo})\widehat{q}_o}{(1-D_{lo})\widehat{q}_o + D_{lo}(1-\widehat{q}_o)}. \tag{6}$$

From this, it can be inferred that

$$D_{lo} = \frac{\widehat{q}_o(1-q_o)}{(1-q_o)\widehat{q}_o + q_o(1-\widehat{q}_o)}. \tag{7}$$

The AF at $i$ ($q_i$) will depend not only on its genetic distance from $l$ ($D_{il}$) and $o$ ($D_{io}$) but also on where it is positioned with respect to the two loci as its relative positions affect the recombinant allele combinations (fig. 3). First, we will consider no crossover interference. When $l$ is between $i$ and $o$, that is, $\mathbf{H}(D_{io}) = \mathbf{H}(D_{il}) + \mathbf{H}(D_{lo})$ (fig. 4A),

$$q_i = \frac{(1-D_{il})(1-D_{lo})\widehat{q}_o + (1-D_{il})D_{lo}(1-\widehat{q}_o)}{(1-D_{il})(1-D_{lo})\widehat{q}_o + (1-D_{il})D_{lo}(1-\widehat{q}_o) + D_{il}(1-D_{lo})\widehat{q}_o + D_{il}D_{lo}(1-\widehat{q}_o)}.$$

This simplifies to

$$q_i = 1 - D_{il}, \tag{8}$$

which indicates that as long as $s_l = 1$, $s_o$ will have no effect on the AF on the distal side of $l$. When $i$ is between $l$ and $o$, that is, $\mathbf{H}(D_{lo}) = \mathbf{H}(D_{il}) + \mathbf{H}(D_{io})$ (fig. 3B),

$$q_i = \frac{(1-D_{il})(1-D_{io})\widehat{q}_o + (1-D_{il})D_{io}(1-\widehat{q}_o)}{(1-D_{il})(1-D_{io})\widehat{q}_o + (1-D_{il})D_{io}(1-\widehat{q}_o) + D_{il}(1-D_{io})(1-\widehat{q}_o) + D_{il}(D_{io})(\widehat{q}_o)}.$$

When $o$ is between $l$ and $i$, that is, $\mathbf{H}(D_{il}) = \mathbf{H}(D_{lo}) + \mathbf{H}(D_{io})$ (fig. 3C),

$$q_i = \frac{(1-D_{io})\widehat{q}_o(1-D_{lo}) + (D_{io})(1-\widehat{q}_o)D_{lo}}{(1-D_{io})\widehat{q}_o(1-D_{lo}) + (D_{io})(1-\widehat{q}_o)D_{lo} + (1-D_{io})(1-\widehat{q}_o)D_{lo} + D_{io}\widehat{q}_o(1-D_{lo})}$$

and reduces to

$$q_i = \frac{(1-D_{io})\widehat{q}_o(1-D_{lo}) + (D_{io})(1-\widehat{q}_o)D_{lo}}{\widehat{q}_o(1-D_{lo}) + (1-\widehat{q}_o)D_{lo}}. \tag{9}$$

To account for interference, the double crossover components are multiplied with the coefficient of coincidence ($2D$ for the Kosambi function) with single crossovers reciprocally increased (see supplementary fig. 1, Supplementary Material online). To illustrate these relationships, we, again, used the recombination rate from the Recombination Rate Calculator (Fiston-Lavier et al. 2010) and varied the fitness differential ($s_o$) at the offsite loci. In both the presence and absence of interference, the AF is more sensitive to negative fitness impact at the secondary loci ($s_o < 0$). At extreme values of $s_o$ (fig. 3D), local maximum and

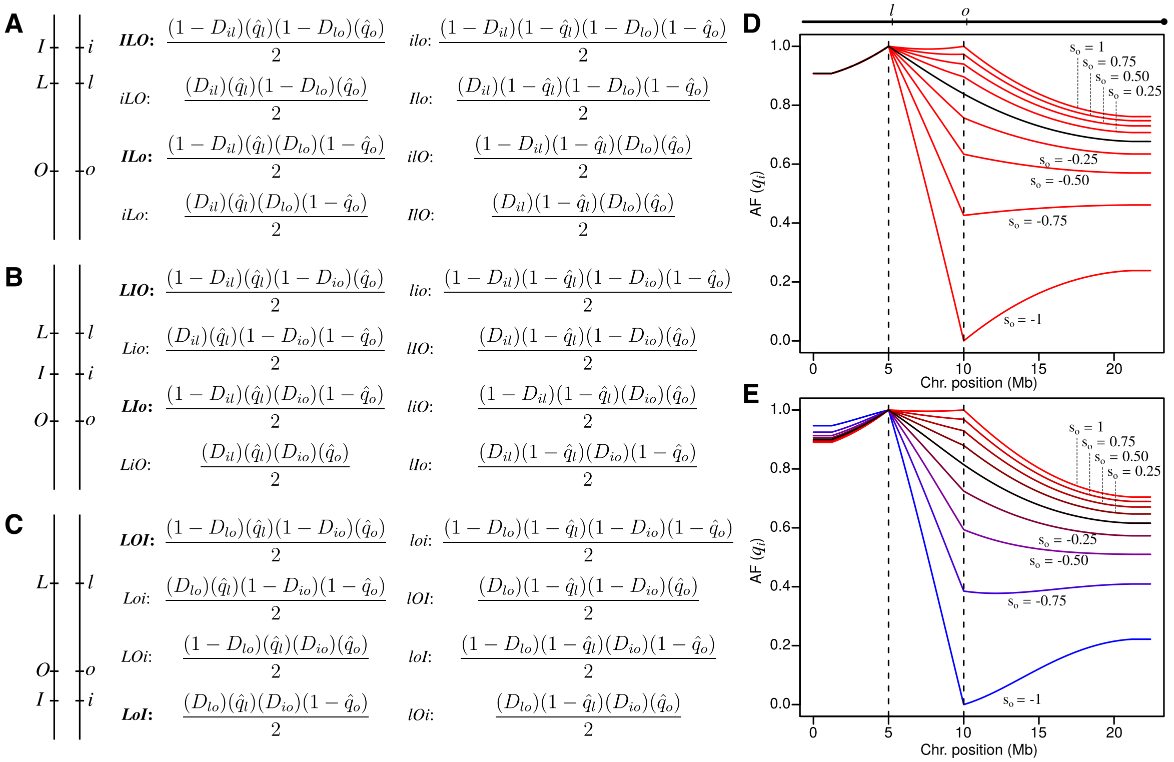

Fig. 3. Effects of a secondary locus on AF decay. In the marker-selected pool, the selected locus ($l$), the offsite locus that affects viability ($o$), and any position along the chromosome ($i$) effectively create a three-point cross. Upper- and lowercase letters of these loci differentiate the homologous alleles. ($A$–$C$) Depending on the relative position of the three loci along the chromosome (schematics on the left), different allele combinations are produced at different frequencies (right). The frequencies of all possible allelic combinations are described by the equations which take into account of the fitness of the selected and offsite loci ($\hat{q}_l$ and $\hat{q}_o$, respectively). When $\hat{q}_l = 1$, only the allelic combinations on the left column are incorporated in the marker-selected pool and allelic combinations on the right will all equal 0. The combinations in bold are those that contain the uppercase I allele. Equations (11)–(14) are derived from dividing the sum of the allelic combinations in bold by the sum of the allelic combinations on the left for each three-point cross. The frequencies of the different allelic combinations assume no interference. ($D$) AF modulation given different values of $s_o$ based on the X chromosome recombination rate, Haldane's transformation, selected locus at 5 Mb, and secondary locus at 10 Mb. Black curve represents $s_o = 0$ (no secondary offsite viability effect). ($E$) Same as ($D$), but with Kosambi's transformation and taking into account of interference with $C = 2D$. Colors are added to differentiate between the curves with different values of $s_o$. Formula for how interference is incorporated into the AFs can be found in supplementary figure 1, Supplementary Material online.

minimums can be observed and identified, but at intermediate levels, the AF curves do not show obvious "kinks," suggesting that moderate loci need to be determined separately (see below). Unlike the curves in the absence of crossover interference, $s_o$ modulates the AF across the selected locus (fig. 3$E$). Unexpectedly, at higher $s_o$, the AF is lower on the other side of $l$. The reason for this is that a positive $s_o$ ensures more individuals lacking crossovers between $l$ and $o$, which, in the presence of interference, results in increased probability of crossovers on the other side of $l$ and therefore faster AF decline.

Given these relationships, it is then possible to solve for $D_{il}$, $D_{lo}$, and $D_{io}$ when $s_o$ and $q_i$ are known. Although $D_{lo}$ can be deduced based on equation (7), we can solve for $D_{il}$ computationally using a root-finding algorithm since $D_{il}$ cannot be easily isolated in the equations (see Materials and Methods).

## Nonlinear Nonparametric Curve Fitting Robustly Estimates AF Decay in Short-Read Whole-Genome Sequences

Because recombinant fraction and crossovers are necessarily sampled from a finite pool of individuals, larger pools will produce more representative and accurate estimates. With marker-selected pools, pool sizes can be increased with no increase to sequencing cost. To evaluate how sampling error affects recombinant fraction in marker-selected pools, we first considered pools where the AF is known exactly across the chromosome. Simulating $w^-$ selection pools of different sizes ($n = 500$, 2,000, and 5,000), we generated pools of individuals with crossovers sampled based on the X chromosome recombination rate (fig. 4$A$, see Materials and Methods). Across all sizes, variability increases as recombinant fraction increases. This increase is partly due to the inherent variance with (binomial) sampling which is highest when $P = 0.5$. It also

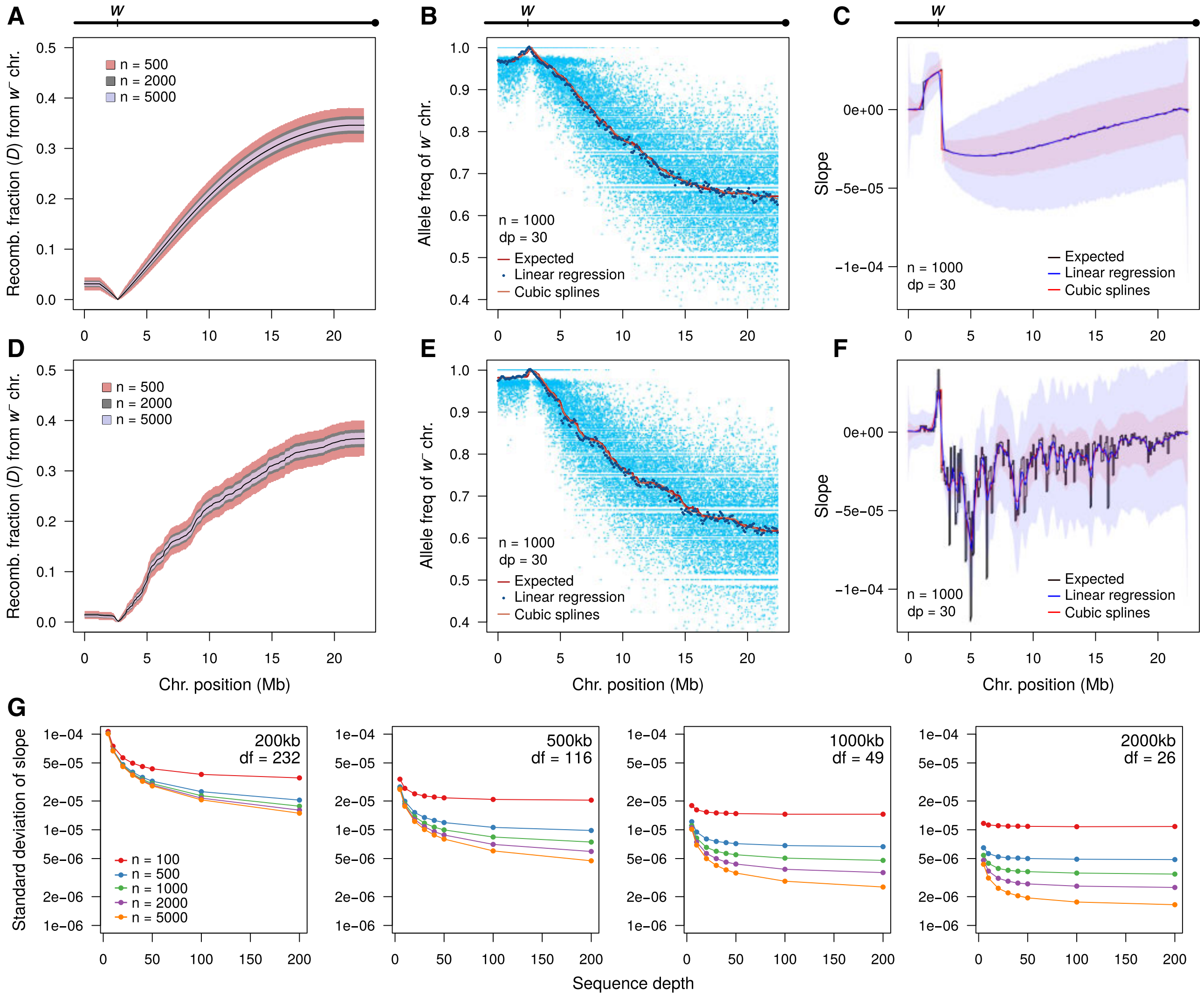

**Fig. 4.** In silico simulation of AF in sequenced pools. (*A*) Simulation of AF attenuation around white in marker-selected pools of 500, 2,000, and 5,000 males. Black line depicts the expected AF. Colored areas demarcate the 95% confidence intervals for the different pool sizes after 20 000 trials, given that AF can be exactly estimated. (*B*) One example of AF distribution from simulated read counts of a pool of 1,000 $w^-$ males. With SNP density of one site per kb, the allele-specific read counts across the chromosome is randomly sampled to simulate fly collection, library preparation, and sequencing to a read depth of 30; the AF at each site is plotted (light blue dots) to illustrate the extent of noise from site to site. Site-specific AF chromosome-wide is either binned in overlapping 500-kb sliding windows followed by linear regression fit (dark blue dots) or fitted with cubic splines anchored at the selected w locus (orange line). Red line depicts the true AF of the simulated pool. (*C*) The slope of the AF decay approximates the recombination rate. The left and right of the selected locus is expected to be increasing (positive slope) and decreasing (negative slope), respectively. The slope of the linear regressions (blue) and cubic spline fit (red) across the chromosome, and the colored areas demarcate the 95% confidence interval of the slope. The expected slope is in black. (*D–F*) Same as (*A–C*), but with the heterogeneous recombination rate inferred by Comeron et al. (2012). (*G*) Summary of the effects of sequence depth (*X*-axes), pools size (colors), and resolution (panels) on the standard deviation of the slope (*Y*-axes). For each combination of parameters, 5,000 simulations were conducted using the Comeron et al. (2012) recombination rate, from which the standard deviation of the slope is derived at each position after curve fitting; the standard deviation averaged across the entire chromosome is plotted.

reflects the increasing probability of double crossovers which results in nonmonotonic decay. Expectedly, these sources of noise decrease with larger pools; at their highest, the 95% confidence intervals are 0.034, 0.0106, and 0.0067 when 500, 2,000, and 5,000 individuals are pooled, respectively (fig. 4*A*).

To evaluate whether short-read whole-genome sequencing can sensitively capture the AF signature in marker-selected pools, we simulated various average sequencing depths with a conservative single-nucleotide polymorphism (SNP) density of one differentiating site per 1 kb (see Materials and Methods). Within a reasonable range of sequencing depth of 10×–100×, allele counts at individual sites are too variable to provide meaningful AF estimates (fig. 4*B* and supplementary fig. 2*A*, Supplementary Material online). Since neighboring sites are expected to have negligible differences in their AFs, their counts and frequencies can be aggregated as if they are independent in sliding windows to better approximate the AF (Wei et al. 2017). However, instead of assuming all sites in a window have the same AF, we first estimated AF using linear regression in overlapping sliding

**Table 1.** Genetic Distance Estimates between Double Markers.

| Parents | | BC1 Offspring Genotypes and Counts | | | | Recombinant Fraction | | | Genetic Distance (cM)[a] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Wild-Type Strain (F) | Marker Strain (M) | Sex | Punnett Squares | | | Overall[b] | Pool Specific[c] | From AF Estimates | Overall | Pool Specific | From AF Estimates |
| | | | | *e+* | *e−* | | | | | | |
| Canton-S | *se′, e′* | F | *se+* | 726 | 226 | 0.2384 | 0.2524 | 0.2414 | 25.95 | 27.78 | 26.33 |
| | | | *se−* | *212* | *628* | | | | | | |
| | | M | se+ | 657 | 183 | | 0.2476 | 0.2352 | | 27.14 | 25.53 |
| | | | *se−* | *178* | *541* | | | | | | |
| Canton-S | *se′, e′* | F | *se+* | 610 | 177 | 0.2370 | 0.2564 | 0.2538 | 25.76 | 28.33 | 27.98 |
| | | | *se−* | *179* | *519* | | | | | | |
| | | M | *se+* | 621 | 151 | | 0.2776 | 0.2367 | | 31.30 | 25.73 |
| | | | *se−* | *191* | *497* | | | | | | |
| Canton-S | *e−, gl−* | F | *gl+* | 534 | 30 | 0.0887 | 0.1339 | 0.1370 | 8.96 | 13.73 | 14.05 |
| | | | *gl−* | 58 | *375* | | | | | | |
| | | M | *gl+* | 485 | *41* | | 0.0921 | 0.1047 | | 9.32 | 10.62 |
| | | | *gl−* | 46 | *404* | | | | | | |
| Canton-S | *e−, gl−* | F | *gl+* | 492 | 54 | 0.0977 | 0.1053 | 0.1155 | 9.89 | 10.69 | 11.76 |
| | | | *gl−* | *40* | *340* | | | | | | |
| | | M | *gl+* | 438 | *45* | | 0.1122 | 0.1256 | | 11.42 | 12.83 |
| | | | *gl−* | 37 | 356 | | | | | | |
| DGRP-360 | *e−, gl−* | F | *gl+* | 1,120 | 107 | 0.0885 | 0.0930 | 0.0791 | 8.95 | 9.41 | 7.97 |
| | | | *gl−* | *102* | 995 | | | | | | |
| | | M | *gl+* | 1,189 | 111 | | 0.0891 | 0.0745 | | 9.01 | 7.51 |
| | | | *gl−* | *113* | *1,155* | | | | | | |
| DGRP-315[d] | *e− gl−* | F | *gl+* | 892 | *89* | 0.0940 | 0.0907 | 0.0966 | 9.52 | 9.17 | 9.78 |
| | | | *gl−* | 89 | *818* | | 0.0981 | 0.1000 | | 9.94 | 10.14 |
| | | M | *gl+* | *1,003* | *113* | | 0.1013 | 0.1165 | | 10.27 | 11.87 |
| | | | *gl−* | *101* | *1,064* | | 0.0867 | 0.0944 | | 8.76 | 9.55 |

Note.—Tracks of different colors represent different crosses. In each row, cells in italics are counts of selected genotypes that were pooled and sequenced.

[a] After Kosambi transformation.

[b] Sum of all recombinants divided by sum of all individuals in the cross.

[c] Recombinant individuals in the pool divided by total number of individuals in the pool.

[d] Individuals in this cross are genotyped and pooled into either $gl^-$ or $gl^+$ pools.

windows (see Materials and Methods). We find that 500-kb overlapping windows in 100-kb increments can roughly recapitulate the expected AF (fig. 4*B*) but create high variance in the slope, which approximates the recombination rate (fig. 4*C*). Second, rather than binning the genome into windows, we implemented a nonparametric and nonlinear curve-fitting strategy using either local regressions (LOESS) (Cleveland 1979) or cubic splines (Perperoglou et al. 2019) (fig. 4*B*). We find this to be more flexible and robust, as it allows for nonlinear fit, is more robust to noise (fig. 4*B*), and estimates the slope of the AF decay with less error than the window-based linear regressions (fig. 4*C*).

To further determine whether this strategy can capture realistic recombination rate distributions which can be highly heterogeneous, we simulated pools using the *D. melanogaster* X chromosome recombination rate inferred from crossover distributions by Comeron et al. (2012) (fig. 4*D–F*). In particular, we were interested as to whether our curve-fitting strategy can approximate the fluctuating slope (fig. 4*F*). As with most fitting methods, there is a tradeoff between resolution and accuracy. We fitted the AF from simulations with different numbers of individuals and sequencing depths at four different resolutions equivalent to 200 kb, 500 kb, 1 Mb, and 2 Mb and inferred the slopes of the fitted curves (fig. 4*G* and supplementary fig. 3, Supplementary Material online). Expectedly, increasing the numbers of individuals, sequencing depth and window sizes produces less variable slope estimates. But between the resolution of 1 and 2 Mb, the gain in accuracy becomes marginal. We note that because our simulation uses a conservatively low SNP density (1 per 1,000 bp), the extent of noise at different resolutions is likely overestimated. Indeed, simulations with a realistic SNP distribution (inferred from inbred *D. melanogaster* strains DGRP 315 and mutant strain *glass ebony*, see below and table 1) which has on average more than three times the SNP density, we find reduced noise across all resolutions (supplementary figs. 4 and 5, Supplementary Material online).

## AF from Whole-Genome Sequencing of Marker-Selected Pools Closely Estimates Recombinant Fraction

To empirically test the efficacy of this approach, we set up recombinant back crosses pairing two double recessive marker strains, *sepia ebony* ($se^-$ $e^-$) and *glass ebony* ($gl^-$ $e^-$), with three wild-derived inbred strains (Canton-S, DGRP-315, and DGRP-360) (table 1). We first sexed and tallied the number of recombinant and nonrecombinant individuals to infer the genetic distance between the two linked markers, then we pooled the flies based on the presence of one of the two markers followed by bulk DNA extraction and whole-genome sequencing to an average of 25.45× coverage (supplementary table 1, Supplementary Material online). Thus, we were able to compare the genetic distance estimates between the two loci based on the de facto method of fly scoring with that from AF attenuation in the marker-selected pools within the same recombinant backcrosses (table 1). Canton-S was separately crossed to the two double marker strains in duplicates; in the crosses to *gl*$^-$ *e*$^-$, we selected and pooled by *gl*$^-$ in females but by *e*$^-$ in males with the expectation that the two pools should yield identical recombination rates. For DGRP-315 and DGRP-360, we crossed each of the two strains only to *gl*$^-$ *e*$^-$ and sexed and pooled by *gl*$^-$. However, with the DGRP-315 cross, in addition to the *gl*$^-$ pool (positive marker selection), we also pooled the *gl*$^+$ individuals (negative marker selection) from the same cross, again with the expectation that the two pools should yield identical recombination rates, despite differences in AF. Allele-specific read counts were determined only at informative SNP sites where the parental strains are homozygous for different nucleotides (see Materials and Methods) (supplementary table 2 and fig. 6, Supplementary Material online). Sites with segregating variants in the parental strains are removed (supplementary table 3 and fig. 7, Supplementary Material online). We then estimated the AF using cubic splines as described above (fig. 5*A*–*D* and supplementary fig. 8, Supplementary Material online). We elected to use a broad resolution of 1 Mb, as the number of individuals and SNP density are low in the Canton-S crosses.

We find that the recombinant fractions and genetic distances estimated from marker-selected pools are highly similar to those from fly scoring with no significant differences ($P = 0.9515$, paired Wilcoxon ranked sum test) (table 1); with the exception of one pool (Canton-S × *se*$^-$ *e*$^-$), estimates are within 1.5% of each other which is comparable to the variability between replicate fly count data as well as different pools from the same cross. Extensively, this demonstrates that the recombinant fraction can be inferred between the selected locus and any site along the chromosome until reaching the Mendelian ratio (fig. 5*A*–*D* and supplementary fig. 8, Supplementary Material online). As expected, when individuals homozygous for the markers are pooled, the AFs of the chromosome carrying the marker peaks at the selected locus with AF of 1 and decrease distally (fig. 5*A* and *C* and supplementary fig. 8, Supplementary Material online). Consistently, for DGRP315 × *gl*$^-$ *e*$^-$ cross, the AF of the negative marker selection pool (*gl*$^-$/*gl*$^+$) expectedly dips to 0.5 at *gl* and attenuates upward (fig. 5*D*). In all these crosses, the AF stays at a near constant level across the centromere and pericentromeres of 3L and 3R, fitting the expectation of no recombination across the region. When converted to recombinant fractions, both the positive (*gl*$^-$/*gl*$^-$) and negative selection (*gl*$^-$/*gl*$^+$) pools yield highly similar results (supplementary fig. 9, Supplementary Material online). Across replicates, and different pools from the same cross, the recombinant fraction appears to be most variable around the pericentric region (fig. 5*E*–*G*), resulting from the reduced SNP density and poor mapping quality due to increased repeat content (supplementary fig. 10, Supplementary Material online). However, the recombinant fractions converge outside the pericentromere, producing highly robust estimates (fig. 5*E*–*G*). Notably, the Canton-S × *gl*$^-$ *e*$^-$ crosses have the highest variance; this is due to the fact that the Canton-S strain used had unexpectedly high levels of heterozygosity which reduces the number of usable informative sites (supplementary table 3 and figs. 6 and 7, Supplementary Material online).

To correct for potential offsite viability effects, we reasoned that any such locus will cause AF to deviate from Mendelian ratios in pools with no marker selection (Wei et al. 2017). To emulate this, we summed the AF from the positive and negative marker selection pools for the DGRP315 × *gl*$^-$ *e*$^-$ crosses where we sequenced both positive and negative marker-selected pools. Using this method, we find a very minor AF deviation at *e*$^-$, which is equivalent to $s = -0.015$ (supplementary fig. 11*A*, Supplementary Material online). Correcting for this resulted in negligible changes in the recombinant fraction (supplementary fig. 11*B*, Supplementary Material online).

Given chromosome-wide recombinant fractions (*D*), the genetic distance (*d*) can then be estimated by transformation with mapping functions (fig. 5*E*–*G*, dashed lines). However, when the recombinant fraction approaches 0.5, the genetic distance estimate approaches infinity. In practice, the genetic distance should, therefore, be limited to 50 cM which entails that two loci are effectively genetically unlinked (fig. 5*E*–*G*); this equates to recombinant fractions of <0.381 and 0.316 with the Kosambi and Haldane functions, respectively. Using the Kosambi transformation, for DGRP-315 and DGRP-360 which were under glass selection, we were able to infer recombination rate for 86.2% and 89.5% of Chr. 3, respectively (fig. 6*A*). The Canton-S pools were selected on both glass and ebony and the composite of the two covered 85.2% of the chromosome.

## Lack of AF Decay Captures Crossover Suppression within Inversions

Interestingly, unlike the crosses with the *gl*$^-$ *e*$^-$ double marker strain, the Canton-S × *se*$^-$ *e*$^-$ crosses produced AF curves that do not have a clear peak flanked by distal decay (fig. 5*B*). The AF on Chr. 3L where the selected locus *se* resides shows minimal decay with elevated level of AF extending into 3R, where it begins to show clear attenuation. Based on the *e*$^-$ selected pool in the Canton-S × *gl*$^-$ *e*$^-$ crosses, the recombinant fraction between *e* and *se* is 0.368 equating to 47.0 cM using Kosambi's map function (fig. 5*E*, blue line), but the estimates from both fly scoring and AF decay using the *se*$^-$ *e*$^-$ marker strain indicate a substantially lower recombinant fraction of 0.238 equating to 25.8 cM (table 1 and fig. 5*F*). Suspecting that the shorter genetic distance resulted from structural rearrangements on the *se*$^-$ *e*$^-$ chromosome arm causing suppression of crossovers and/or lethal rearrangements in recombinants, we used the structural variant discovery software Lumpy (Layer et al. 2014) to infer the presence of inversions. Indeed, we identified two large inversions on 3L: one from 0.25 to 7.22 Mb and another pericentric inversion from 7.81 Mb of 3L to 2.08 Mb of 3R (fig. 5*B* and *F*), confirming

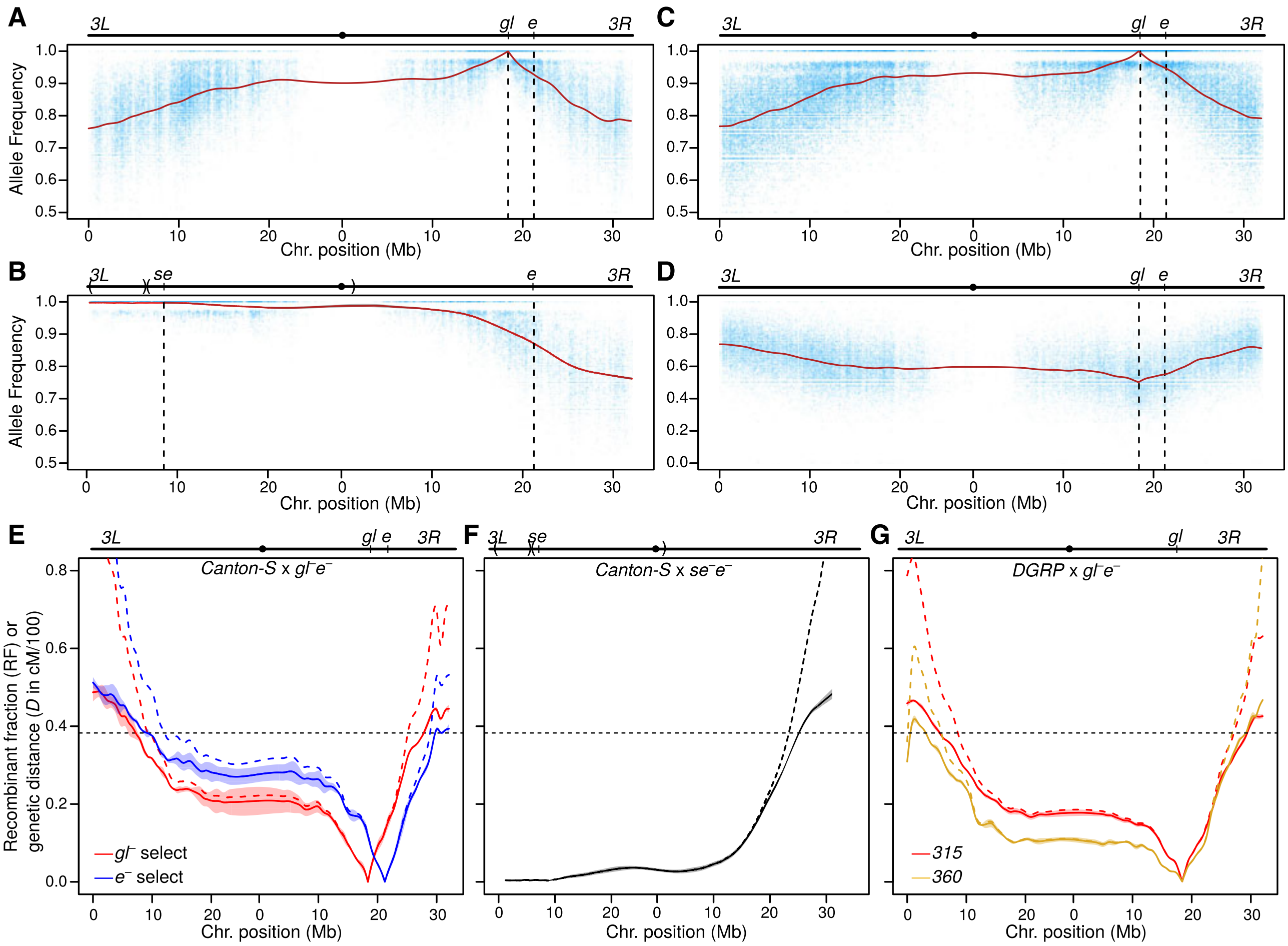

Fig. 5. Recombinant fraction and genetic distance inferred from marker-selected pools for Chr. 3. The per site (blue) and inferred (red line) AFs across the chromosome are displayed for four pools: (*A*) Canton-S × *gl*$^-$ *e*$^-$ with *gl*$^-$ selection (table 1, row 6), (*B*) Canton-S × *se*$^-$ *e*$^-$ with *se*$^-$selection (table 1, row 1), (*C*) DGRP-315 × *gl*$^-$ *e*$^-$ with *gl*$^-$ selection (table 1, row 12), and (*D*) DGRP-315 × *gl*$^-$ *e*$^-$ with *gl*$^+$ selection (table 1, row 13). For all marker-selected pools, see supplementary figure 6, Supplementary Material online. Note that (*D*) has a different *Y*-axis scale. The positions of the double markers are marked in the chromosome schematics above and with vertical dotted lines in the plots. For the Canton-S × *se*$^-$ *e*$^-$ cross, large inversions are marked in chromosome schematic above with parentheses. (*E*–*G*) Averaged recombinant fraction (solid curve) and genetic distance (dotted curve) from the selected loci. Interval around the recombinant fraction estimates represent the standard error across the replicates. Horizontal dotted lines represent recombinant fraction equivalent of 50 cM after Kosambi transformation. (*E*) Canton-S × *gl*$^-$ *e*$^-$ crosses when *gl*$^-$ (red) or *e*$^-$ (blue) are selected. Note here that *gl*$^-$ pools were the female pools and *e*$^-$ pools were the male pools. (*F*) Canton-S × *se*$^-$ *e*$^-$ crosses with *e*$^-$ selection. (*G*) DGRP-315 (red) or DGRP-360 (yellow) × *gl*$^-$ *e*$^-$ crosses. For the DGRP-315 cross, values are averaged across *gl*$^+$ and *gl*$^-$ selection pools, whereas only *gl*$^-$ pools were sequenced for DGRP-315.

our suspicion that inversions are interfering with recombination in this region. Interestingly, we still observe minor reduction in AF across this region. This likely reflects rare double (or even number) crossover events which resolve lethal recombinant rearrangements caused by single (or odd number) crossover events within the inversion (Hughes et al. 2018).

## Heterogeneity and Natural Variation in Recombination Rate

Since recombination rate is genetic distance per physical distance, for regions to the left and right of the selected locus, the recombination rate is the negative and positive slope of the genetic distance curve, respectively. We converted the genetic distance to recombination rate for all the *gl*$^-$ *e*$^-$ crosses by taking the slope of *d* in 10-kb windows within the ±50 cM range (fig. 6A). In all three lines, the recombination rate is minimal across the highly repetitive regions and gradually increases away from the centromere, as expected. Our estimates broadly follow previous estimates based on crossover breakpoint distribution (Comeron et al. 2012) (fig. 6A, gray). Both estimates of the DGRP lines are highly and significantly correlated with each other (supplementary fig. 12, Supplementary Material online) and that of Comeron et al. (2012) ($P < 2.2 \times 10^{-16}$), as well as when it is smoothed into an equivalent resolution (fig. 6B and C). The significant positive correlations, similarly, hold at the resolution of 500 kb; but at 200 kb, the recombination rate estimates become too unreliable due to overfitting (supplementary fig. 13, Supplementary Material online). Much of the correlation appears to be driven by regions with low recombination rate; after removal of these sites the correlations, although still significant, are substantially reduced (supplementary fig. 14, Supplementary Material

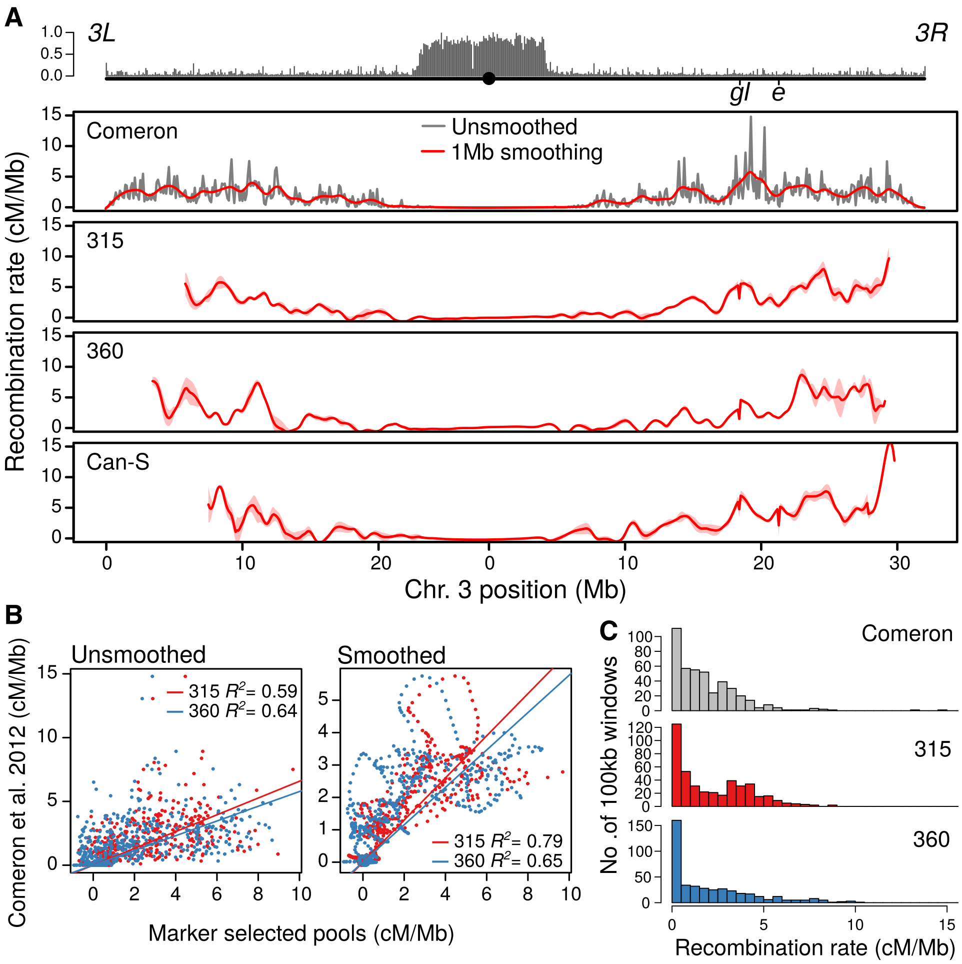

Fig. 6. Recombination rate estimates in wild-derived strains. (*A*) Chr. 3 recombination rate estimates from Comeron et al. (2012), and marker-selected pools are depicted. Repeat content (%) and Chr. 3 schematic is plotted above. For the estimates from Comeron et al. (2012), the rates are plotted as the true rates (gray) and the rates after 1 Mb smoothing (red). For the marker-selected pools (DGRP-315, DGRP-360, and Canton-S), rate estimates are restricted to ±0.381 recombinant fraction (equating to 50 cM after Kosambi's transformation) from the marker. Area around the line demarcates the standard error estimated across the different pools. (*B*) Correlation of recombination rates of DGRP-315 (red) and DGRP-360 (blue) with unsmoothed (left) and 1-Mb-smoothed estimates (right) from Comeron et al. (2012). Each point is the recombination rate estimate of a 100-kb window. The red and blue lines are the least square regressions. (*C*) Distribution of the recombination rate of estimates from Comeron et al. (2012) (gray) and of DGRP-315 (red) and DGRP-360 (blue).

online). Discrepancies are likely partially due to strain-specific rate variations (see below) and differences in rearing conditions as our crosses were reared in 25 °C versus 21 °C in that of Comeron et al. (2012).

Interestingly, all three lines show similar regions with elevated rates each of which spans multiple megabases and are interrupted by valleys (fig. 6*A*). Since these regions are also elevated in the estimates by Comeron et al. (fig. 6*A*), they are unlikely to be artifacts of our method, or the strains used. These broad megabase-sized blocks of elevated recombination rates and narrow valleys of low recombination are likely a general feature of the recombination landscape of the chromosome.

Despite global similarities in the recombination landscape among the different lines, recombination rate is highly variable between them. From 5.5 Mb of Chr. 3L to 29.1 Mb of Chr. 3R where recombination rate can be determined for both DGRP-315 and DGRP-360, these two lines and the Comeron et al. (2012) estimates show significantly different distribution of recombination rates ($P < 0.001$, paired Wilcoxon ranked sum test; fig. 6C). DGRP-315 has more regions with intermediate recombination rates (3–5 cM/Mb), whereas DGRP-360 is heavily skewed by a large fraction of regions with a rate close to 0 but has a longer tail of regions with high recombination. The large fraction of low recombining windows caused the DGRP-360 strain to have a shorter genetic map to the left of the selected locus (fig. 5G and supplementary fig. 15, Supplementary Material online). However, to the right of the selected locus, the genetic distance and recombination rates are comparable between the two lines (fig. 5G and supplementary fig. 15, Supplementary Material online). This suggests that the lower recombination rate and shorter distance in the DGRP-360 line is not due to global depression of crossovers, but perhaps many local modifiers to the left of glass, possibly in the form of small scale inversions and TE insertions.

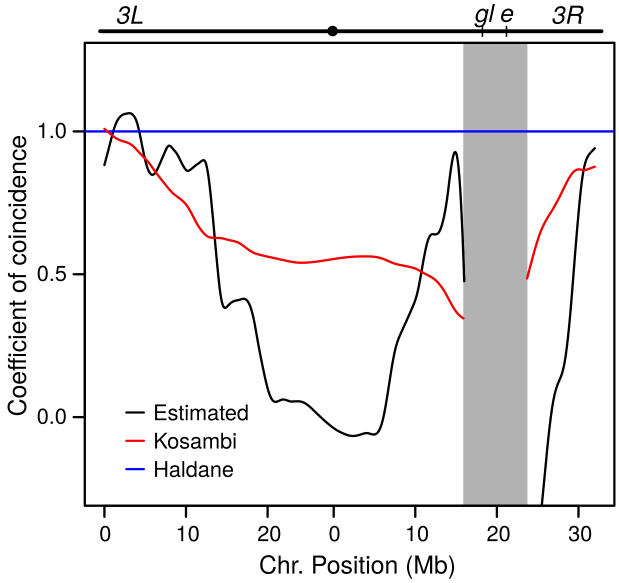

Fig. 7. Estimating coefficient of coincidence from reciprocal marker selection pools with double markers. In the Canton-S × $gl^-$ $e^-$ crosses, males were selected by $e^-$ and females were selected by $gl^-$. The chromosome-wide recombinant fractions from these two selected pools can be used to estimate the coefficient of coincidence across the chromosomes with respect to these two loci (black line). Estimates near the selected locus (within 2.5 Mb) are excluded as the values approach their limits. For reference, the coefficient of coincidence based on the Haldane and Kosambi mapping functions is in blue and red, respectively.

### Inferring Chromosome-Wide Crossover Interference from Reciprocal Selection of Double Markers

Crossover interference, usually expressed as $1 - C$ where $C$ is the coefficient of coincidence (Muller 1916), is typically estimated from three-point crosses where double recombinants can be scored. $C$ is the fraction of observed double crossovers over expected, and, given three loci ABC in that order, is expressed as

$$D_{\mathrm{AC}} = D_{\mathrm{AB}} + D_{\mathrm{BC}} - 2CD_{\mathrm{AB}}D_{\mathrm{BC}}. \qquad (10)$$

Therefore, if the recombinant fraction between three loci is known, $C$ can be inferred by

$$C = \frac{D_{\mathrm{A,B}} + D_{\mathrm{B,C}} - D_{\mathrm{A,C}}}{2D_{\mathrm{A,B}/}D_{\mathrm{B,C}}}. \qquad (11)$$

In our Canton-S × $gl^-$ $e^-$ crosses, we sequenced pools of both $gl^-$ and $e^-$ selection (table 1), generating chromosome-wide recombinant fractions with respect to these two loci. With these two curves, we can estimate the coefficient of coincidence, as for any site $i$, we have the genetic fractions between $gl$ and $i$ ($D_{i,gl}$), $e$ and $i$ ($D_{i,e}$), and $gl$ and $e$ ($D_{gl,e}$). Therefore, to the left of $gl$, between the two, and right of $e$, we can solve for $C$ with the following, respectively:

$$C = \frac{D_{i,gl} + D_{gl,e} - D_{i,e}}{2D_{i,gl}D_{gl,e}}, \qquad (12)$$

$$C = \frac{D_{i,gl} + D_{i,e} - D_{gl,e}}{2D_{i,gl}D_{i,e}}, \qquad (13)$$

and

$$C = \frac{D_{i,e} + D_{gl,e} - D_{i,gl}}{2D_{i,e}D_{gl,e}}. \qquad (14)$$

For this proof-of-principle analysis, we used the broad window size of 2 Mb to minimize error in estimates of $D$, since these crosses have a low density of informative sites. Because of the denominator, regions close to loci with small $D$ will yield nonsensical results. We, thus, only estimate the interference beyond ±2.5 Mb from the markers, which also precludes estimating between $gl$ and $e$, as they are only 2.8 Mb apart (fig. 7). To the right of $e$, $C$ increases distally toward the 3R telomere. To the left of $gl$, $C$ drops sharply and remains low around the pericentomere, indicating high interference. Across the pericentromere of 3L, $C$ increases rapidly toward the 3L telomere. In comparison to the $C$ assumed by the Kosambi function, the estimated $C$ increases more precipitously, reaching 1 (i.e., no interference) before that of Kosambi's function. Historically, the centromere has been assumed to act as a barrier to interference, disrupting the propagation of interference across it (Mather 1939; Colombo and Jones 1997). Similar to later studies (Colombo and Jones 1997), our chromosome-wide estimates of $C$ are inconsistent with this assumption and show that interference is instead high around the centromere (Hultén 2011), thus further reducing the already low probability of crossovers within the pericentromere.

## Discussion

### Broad-Range Recombination Rate Estimation Using AF in Marker-Selected Pools

We demonstrated theoretically and empirically that AF decay around a locus with selection differential in a recombinant backcross can be used to infer near chromosome-wide recombination rates. Instead of genotyping individuals for recombinant breakpoints, this approach relies on pooled sequencing of large numbers of marker-selected individuals from recombinant backcrosses. Since the AF surrounding the selected locus decays proportionally with the recombinant fraction, the genetic distance can be determined using mapping functions, and the slope will then approximate the recombination rate. Since each pool can provide broad-range recombination rate estimates, this method substantially reduces the number of library preparations needed for chromosome-wide estimates. Previously, hundreds to thousands of individuals needed to be sequenced or genotyped to capture a comprehensive spread of recombinant breakpoints from which chromosome-wide recombination rates are inferred, but here we used as little as two libraries for near chromosome-wide estimates. Recently, crossover detection has been made possible in bulk with "linked read sequencing," where large molecular weight DNA molecules are effectively barcoded in high throughput by 10× Genomics (Sun et al. 2019). Nevertheless, the sequencing depth required for this approach is still substantially greater than the 30× used here. Not accounting for fly upkeep, each library of a marker-selected pool costs <$400 to generate and sequence. Using this approach, we inferred the recombination rate of Chr. 3 of three wild-derived lines and demonstrated that how it can be used to infer crossover interference.

Previously, Singh et al. (2013) also used AF to estimate fine-scale crossover frequency; recombinants between *garnet* and *scalloped* (two phenotypic maker on the *D. melanogaster* X chromosome) that carry either one of the two markers were pooled to estimate the fine-scale recombination rate between the two loci. The method presented here differs from Singh et al. (2013) in that recombinants are not explicitly targeted for selection, and thus the AF decay is directly reflecting the recombinant fraction (eqs. 1–3). When only recombinants between two loci are selected, the AF decay is effectively restricted to estimates of recombinant fraction between the two loci. As demonstrated by Gilliland (2015), the approach used by Singh et al. (2013) to measure AF and recombination rate produces rates that are strongly anticorrelated with the heterogeneous window sizes used for estimation. As shown by our simulations and pointed out by Gilliland (2015), the variability of read counts at individual sites is too large for fine-scale estimates. Instead, we use curve-fitting methods leveraging all informative sites across the chromosome to avoid the inherent noise associated with read counts in small window sizes.

## Methodological Limitations and Workarounds

Given that crossover breakpoints are not identified in marker-selected pools, recombination rate is estimated from changes in recombinant fraction, which is itself inferred from AF. The accuracy of AF estimation is dependent on the sequence coverage, SNP distribution, and the curve-fitting strategy. However, whether the AF decay (even if perfectly estimated) accurately reflects the recombinant fraction depends on the number of individuals in the pool (see fig. 4*A* and *D*). With smaller pool sizes, sampling error of genotypes leads to greater levels of noise in the AF decay; this is akin to inferring crossover breakpoint distribution from a small number of individuals. Although increasing the number of individuals necessitates increasing the number of DNA extractions, library preparations, and sequencing to infer crossover breakpoints, the number of individuals in marker-selected pools can be easily increased with no increase to sequencing cost, especially for highly fecund invertebrates like worms and flies. Therefore, although the resolution of this method is markedly lower than the rates from crossover breakpoints, it can potentially generate more representative maps with very large numbers of individuals at a fraction of the sequencing cost.

There are several additional limitations to sequencing marker-selected pools, although some of the issues are not specific to this method of estimating recombination rates. First, the conversion from recombinant fraction to genetic distance with mapping functions requires assumptions about crossover interference. The Haldane, Kosambi, and other mapping functions have explicit assumptions about the extent (or lack) of interference, which, accordingly, affect the genetic distance conversion, with higher crossover interference producing shorter genetic maps. On the other hand, we presented an extension of our method to infer crossover interference (fig. 7); by reciprocally selecting for markers in a double marker strain, we were able to estimate the coefficient of coincidence across Chr. 3 with respect to the two markers. Improved understanding of crossover interference in a genomic context will allow for more accurate conversion between the recombinant fraction to genetic distance in future investigations. Alternatively, since the mapping functions have negligible effects for smaller values of recombinant fractions (<0.10), multiple markers (if available) along the chromosome can be used to estimate recombination rate in smaller intervals. Although this significantly increases the number of crosses, the number of libraries generated is still at least one order of magnitude less than sequencing individuals. As a corollary, mapping functions break down when the recombinant fraction approaches 0.5. Here, we explicitly restrict the Kosambi mapping function to a genetic distance of 50 cM on either side of the selected loci, which translates to a total measurable recombinant fraction of 0.762 (0.381 on either side). Again, to extend the genetic map, multiple markers strategically chosen along the chromosome may be needed to encompass the entirety of the chromosome.

Second, this approach is sensitive to the SNP density. Based on simulations, our method is robust even when the SNP density is as low as one site per 1,000 bp. Although intraspecific strain differences are likely higher in flies, polymorphisms are not evenly distributed across the chromosome. The SNP density drops rapidly around the pericentromeric and telomeric regions, which resulted in increased error rates. The decrease in SNP density is particularly problematic in window-based AF estimates, but our usage of nonlinear nonparametric curve fitting alleviates this issue on at least two fronts: The smoothing is conducted based on the number of sites instead of genomic windows and the reduced pericentromeric SNP density for metacentric chromosomes (e.g., Chromosome 3) is flanked by high density on either sides thus producing more robust estimates across the pericentric region. However, we note that reduced SNP density also poses a challenge when inferring breakpoints, since the precise location of haplotype changes will be difficult to pinpoint.

Lastly, we demonstrated that offsite viability effects can modulate the AF decay around the selected locus. Such viability effects can result from alleles that induce lethality or reduce body size, both of which will change the AF in the DNA pool to be sequenced. Lethality is similarly problematic for genotyping and breakpoint inference in individuals, since it changes the number of recoverable recombinants at specific loci, but are typically ignored. We analytically showed how additive offsite viability effects can be accounted for, provided that the extent of the fitness reduction ($s_o$) can be determined. To estimate $s_o$ without additional experiments, we simply summed the AF of both the positive and negative marker selection pools that originated from the same cross. This effectively removes the peak and the effect of the selected locus and the remaining elevations and drops that deviate from Mendelian ratios are then regions with viability effects.

### Applications beyond *Drosophila*, Model Genetic Organisms, and Phenotypic Markers

Although this study focused on *D. melanogaster*, this method is readily applicable to other organisms, particularly for model species with a wealth of phenotypic markers readily available like *Caenorhabditis elegans* (Greenwald 2016) or mice (Singh and Coppola 2014). As discussed above, for species with longer genetic maps, multiple markers are needed to capture recombination rate along the entirety of the chromosome, thus increasing the number of crosses and libraries. Since each marker-selected pool only reflects recombination rate on one chromosome, species with higher chromosome numbers will require more libraries and crosses overall. Notably, markers from different chromosomes can be selected at the same time in any given cross allowing for estimates on different chromosomes in parallel. However, this can introduce potential epistatic effects between markers which may modulate AFs of the chromosomes. As with typical recombinant crosses, the sensitivity of the method depends on the number of BC1 individuals (fig. 3A). Although we recommend over 1,000 individuals to be collected, for organisms where large numbers of offspring are unfeasible, the estimates from marker-selected pools will be similarly underpowered as counting or genotyping the number of recombinant individuals between markers, but with the added benefit of requiring only one marker. For species like mice where pulverizing individuals is impractical, care must be taken when pooling tissues to minimize variation in tissue size and the resulting DNA contribution per individual.

Although the abundance of visible markers make our method particularly suitable for model genetic organisms, visible polymorphisms and mutations can be found across wide ranges of species and taxa enabling application of this method; within the *Drosophila* genus, many species, even those distantly related to *D. melanogaster*, have marker stocks readily available (Clark et al. 2007, and see The National Drosophila Species Stock Center). Similarly, many nonmodel or emerging model organisms including *Apis mellifera* (Schulte et al. 2014), *Anopholes gambiae* (Bernardini et al. 2018), *Bombyx mori* (Yasukochi 1998), *Musca domestica* (McDonald et al. 1975; Meisel et al. 2017), *Gerris buenoi* (Armisén et al. 2018), *Daphnia magna* (Ismail et al. 2018), and *Parhyale hawaiensis* (Ramos et al. 2019) have not only selectable phenotypes permitting this approach but also high quality reference genomes.

Moreover, we showed that the signature AF attenuation enabling estimation of recombination rate is generated at any locus with a fitness differential (eq. 5 and fig. 2C). However, AF attenuation will be easier to estimate given stronger fitness differentials. The closer $s_l$ is to 0, the recombinant fraction will be more difficult to tease apart from the noise introduced by sample pooling and sequencing. Systems with segregating recessive lethals (McCune et al. 2002) and strong meiotic drivers (Fishman and Willis 2005) are therefore prime candidates for this method. Even without a source of lethality or segregation bias, loci with associated traits, even if not fully penetrant, can also be used to generate the fitness differential required. Additionally, advances in gene editing with CRISPR-Cas9 (Russell et al. 2017; Adli 2018) and ease of genome assembly with long-read sequencing (Miller et al. 2018; Bracewell et al. 2019) will continue to increase the catalog of organisms in which marker-selected pools can be applied. Therefore, our theoretical, statistical, and empirical investigations here set the stage for wide application of this cost-effective and scalable method to estimate recombination rate.

## Materials and Methods

### MarSuPial

The analytical and statistical methods described below are implemented in the MarSuPial package found in KW's github page (https://github.com/weikevinhc/Marsupial). It is an R package with tools to analyze and simulate read count data from marker-selected pools.

### Conversion of Recombinant Fraction from Recombination Rate Function

To convert recombination rate function (from Recombination Rate Calculator [Fiston-Lavier et al. 2010]) to genetic distance centered at the selected locus (e.g., w), we integrated the quadratic formula from the selected locus to every position on the chromosome. For example, the genetic distance between w (at position X : 2,684,632) and position *i* in megabses is

$$\int_{2.684632}^{i} \left(-0.03x^2 + 0.6x + 1.15\right) dx.$$

As per Fiston-Lavier et al. (2010), positions <1.22 Mb and >21.21 Mb have rates of 0. To convert genetic distance to recombinant fraction with the Haldane and Kosambi mapping functions, we used their inverse functions:

$$D = -\ln\frac{1-2d}{2}$$

(Haldane 1919) and

$$D = \frac{1}{4}\ln\frac{1+2d}{1-2d}$$

(Kosambi 1943), respectively.

Since the formula from the Recombination Rate Calculator are based on r5 of the *D. melanogaster* genome, we used the r5 instead of r6 coordinates for the genes in the simulations. The remaining analyses were all based on the r6 reference.

### Simulation of Recombinant Fraction and Pooled Sequencing of Marker-Selected Pools

To estimate the variance of recombinant fraction resulting from sampling errors, we first simulated crossover events based on the recombination rate function of Chr. X. For each 1,000-bp window, the probability of a crossover is determined by integrating over the recombination rate of that window:

$$\int_{i}^{i+0.001} \left(-0.03x^2 + 0.6x + 1.15\right) dx.$$

To simulate crossover events for one chromosome, a Bernoulli draw is conducted at each window, with successes denoting crossovers. Starting from the selected locus with an allele state of TRUE, every Bernoulli success causes a change in allele state (represented in Boolean as either TRUE or FALSE) of all subsequent windows. To simulate $n$ individuals, this process is repeated $n$ times. The AFs across the windows are then determined based on the proportion of 0 and 1 allele states in the pool of $n$ individuals. The AF of the one allele equals the recombinant fraction. This process is repeated 20,000 times to determine the error rate of recombinant fraction given different $n$. This simulates crossovers without interference as each crossover is independent. For read count and AF simulations, the AF ($q$) (or $[q+1]/2$ if there is paternal contribution) from the simulated pools (above) is used for two random draws. A Poisson draw is first conducted with the desired read depth as the mean. A second binomial draw is then conducted with probability of $q$ and the Poisson-drawn read depth as the number of trials, to simulate allele-specific read counts at the site.

### Predicting and Smoothing the AF with Linear Regression, LOESS, and Cubic Splines

Predicting and smoothing the AF with linear regression, LOESS, and cubic splines once AF at differentiating sites is determined from read counts (simulated or real) across the chromosome, sites within 500-kb windows are then used for a weighted linear regression with the R function lm() where the weight of each site is the coverage of that site. The AF of the window is then determined for the midpoint of the window using the slope and $y$-intercept of the linear regression. Given the linear regression, AF can also be predicted for any point within the window. AF in 500-kb windows are determined every 100 kb, resulting in overlapping sliding windows. For LOESS and cubic splines fitting, we implemented the R function loess() and smooth.splines() in MarSuPial, respectively, to fit two curves on either side of the selected locus, which prevents smoothing of the expected peak/trough. The AF at each site is weighted by the coverage at the site. To "anchor" the functions, we include an additional point at exactly the selected locus with the expected AF (0, 0.5, or 1, depending on the selection procedure and/or chromosome), with a weight of 1,000,000, ensuring that the curves intersect the expected AF at the selected locus. The predict() function is used on these curve-fitting objects to estimate the AF and standard error for any position across the chromosome. For cubic splines, we used degrees of freedom (df) as a proxy for resolution. Since df − 4 number of knots are evenly spaced across the sites for cubic splines, window size = chromosome size/(df − 4). With cubic splines, the first derivative (slope) of the fitted curves is determined using the predict() function in R.

### Genetic Distance and Recombination Rate from AF

Given the LOESS-predicted AF chromosome wide in 5-kb windows for each cross, we removed the paternal contribution and inferred the recombinant fraction using equation (3). The only exception is the DGRP315 × $gl^{-}$ $e^{-}$ pool where the $gl^{+}$ individuals were selected (negative marker selection); instead, we used the formulae:

$$q_i = \frac{D_i + 1}{2}.$$

To get the genetic distance, we then applied either Haldane's or Kosambi's mapping functions:

$$d = -\frac{1}{2}\ln(1-2D)$$

and

$$d = \frac{1}{4}\ln\frac{1+2D}{1-2D},$$

respectively. Recombination rate ($r$) at position $i$ bp was then derived taking the positive or negative slope between the 5-kb windows:

$$r_i = \frac{d_{i+2,500} - d_{i-2,500}}{5,000},$$

depending on whether $i$ was to the left or right of the selected locus, respectively. Summary statistics for genetic distances and recombination rates (median, standard deviation, etc.) were estimated from replicate and/or different sexed pools from the same cross. For the Canton-S × $gl^{-}$ $e^{-}$ crosses, the reciprocal marker selection produced two different sets of genetic distances from either glass or ebony. However, once converted into recombination rate, the reciprocals were then treated as replicates, since recombination rate is expected to be unaffected by the marker selected. To convert the publically available recombination rate which is in 100-kb windows Comeron et al. (2012) to genetic distance, we multiplied the recombination rate in each window by 100 kb.

### Fly Stocks, Maintenance, and Collection

The Canton-S and double marker lines were ordered from Bloomington Drosophila Stock Center, stock numbers BL64349 (Can-S), BL1669 ($e^{-}$ $se^{-}$), and BL507 ($gl^{-}$ $e^{-}$). Notedly, the source of the elevated heterozygosity in the Canton-S strain is due to stock center contamination, which BDSC has now acknowledged under the strain's listing. DGRP-315 and DGRP-360 are gifts from Dr Yuh Chwen G. Lee. All stocks and crosses were raised on standard molasses food at 25 °C. For the crosses, 4–8-day-old virgin females were mated with the marker strain males and F1 virgins were then collected. For each recombinant cross, over 40 F1 virgins were collected and backcrossed to the marker males in vials of five to eight virgin females to eight to ten males. To avoid overcrowding in the subsequent generation, backcrosses were transferred every 2–3 days to new vials and after adults began to emerge, flies were cleared and scored daily. Every vial is collected for 10 days to ensure that genotypes that may introduce developmental delays will not be underrepresented. Sexed and genotyped flies were maintained in fresh vials for 3–5 days prior to freezing in the −20 °C freezer.

### DNA Extraction and Library Preparation

Frozen flies of the desired genotypes were pooled, pulverized with sterilized mortar and pestle that are chilled in liquid nitrogen, and then transferred to 50-ml falcon tubes. After adding 15 ml of Cell Lysis Solution from Qiagen (Catalog No.158908), samples were incubated at 65 °C for 4 h and vigorously shaken every hour. In total, 75 $\mu$l of ProteinaseK (Catalog No. 19131) was then added and incubated at 55 °C overnight; 200 $\mu$l of EtOH was added into 400 $\mu$l of the sample then passed through the columns from the DNeasy kit (Catalog No. 69506). The columns were then processed in accordance with the kit protocol. DNA for parental lines were extracted from five females using the DNeasy kit. The resulting DNA was fragmented to 550 bp using the Covaris sonicator and libraries were made with the Illumina Truseq DNA Nano kit. Library quality was determined with the Bioanalyzer at the Functional Genomics Laboratory at UC Berkeley and samples were pair-end sequenced using the Illumina HiSeq 4000 machine at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley. Coverage of each cross can be found in supplementary table 1, Supplementary Material online.

### Predicting and Smoothing the AF with Linear Regression, LOESS, and Cubic Splines

Demultiplexed paired-end reads were mapped to the *D. melanogaster* genome r6.12 downloaded from Flybase (Thurmond et al. 2019), using bwa mem (v0.7.15) on default settings (Li and Durbin 2009). Raw reads for the two DGRP strains were downloaded from SRA under SRX006143 for DGRP-315 and SRX155999 for DGRP-360. We removed duplicates using Picard tools (v2.18.14) and merged the parental strains with the crosses using Samtools (v1.5) (Li et al. 2009) to allow them to be genotyped together with GATK HaplotypeCaller (v3.8) (McKenna et al. 2010). By default, HaplotypeCaller only outputs sites where at least one sample has a nonreference variant. This is particularly problematic if samples were genotyped individually, since many of the sites will be deemed as homozygous reference and unreported, particularly around the selected locus in our selected pools where the AF is close to either 0 or 1. By genotyping the crosses together with the parents, we are ensuring that all informative sites are reported, because at least one of the two parents will be homozygous for the nonreference allele at informative sites. To filter for informative sites, we used bcftools (Li 2011) (v1.6) to first isolate the parental strains and retained only SNP sites where the parental strains have different homozygous alleles with both genotype quality of ≥30 (supplementary table 2, Supplementary Material online). In the crosses, all sites other than the informative sites are removed. No genotyping filter is applied on the crosses since many of the sites in the crosses will have intermediate AF that are difficult to genotype. Sites within 100 bp of repeats are removed to avoid copy number variants. The coverage and number of SNP sites after filtering can be found in supplementary table 1, Supplementary Material online. The allele-specific read counts can be determined from the AD field in the vcf files (Danecek et al. 2011).

### Inversion Identification

For structural variant calls, we used the smoove wrapper for Lumpy (v0.2.13) after aligning *e se* to the reference genome (Layer et al. 2014). We identified large structurals variant over 1 Mb within and between Chr. 3L and 3R.

### Removing Offsite Viability Effects with Root-Finding Algorithm

Given the complex equations for the offsite viability effects, we use root-finding algorithms to solve for $D_{il}$ instead of isolating it from the equation. Since $\widehat{q}_o$ and $D_{lo}$ can be predetermined, and $D_{io}$ can be substituted with $D_{il}$ in accordance with equation (9), we are left with formulae with only $D_{il}$ as the variable to solve. We use the root-finding function uniroot.all() from the package rootSolve in R to closely approximate the solution (Soetaert and Herman 2009). Note in some instances, more than one solution is possible, but usually only one is within reasonable range (between 0 and 0.5).

## Data Availability

Intermediate files can be found on Dryad at https://doi.org/10.6078/D10D8F.

## Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

## Acknowledgments

We sincerely thank Dr Emily Landeen (SM), Dr Michael McGurk, and Dr Daniel Barbash for insightful discussions on the method; Dr Ching-Ho Chang, Dr Grace Lee, Dr Andrew Clark, Dr Mohamed Noor, and Dr Daven Presgraves for comments on drafts of the manuscript; Dr Ted Chen for consultation regarding R and LaTeX; Dat Mai for help with Lumpy; and Dr Jeffrey Spence and three additional anonymous reviewers for their insightful reviews. This work is supported by National Institute of General Medical Sciences at the NIH (Grant No. R01GM101255 to D.B.).

## References

Adli M. 2018. The CRISPR tool kit for genome editing and beyond. *Nat Commun*. 9(1):1911.

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

Armisén D, Rajakumar R, Friedrich M, Benoit JB, Robertson HM, Panfilio KA, Ahn S-J, Poelchau MF, Chao H, Dinh H, et al. 2018. The genome of the water strider *Gerris buenoi* reveals expansions of gene repertoires associated with adaptations to life on the water. *BMC Genomics*. 19(1):832.

Barroso GV, Puzović N, Dutheil JY. 2019. Inference of recombination maps from a single pair of genomes and its application to ancient samples. *PLoS Genet*. 15(11):e1008449.

Bernardini F, Haghighat-Khah RE, Galizi R, Hammond AM, Nolan T, Crisanti A. 2018. Molecular tools and genetic markers for the generation of transgenic sexing strains in anopheline mosquitoes. *Parasit Vectors* 11(Suppl 2):660.

Blixt S. 1975. Why didn't Gregor Mendel find linkage? *Nature* 256(5514):206.

Bracewell R, Chatla K, Nalley MJ, Bachtrog D. 2019. Dynamic turnover of centromeres drives karyotype evolution in Drosophila. *Elife* 8:e49002.

Brand CL, Cattani MV, Kingan SB, Landeen EL, Presgraves DC. 2018. Molecular evolution at a meiosis gene mediates species differences in the rate and patterning of recombination. *Curr Biol*. 28(8):1289–1295.e4.

Brush SG. 2002. How theories became knowledge: Morgan's chromosome theory of heredity in America and Britain. *J Hist Biol*. 35(3):471–535.

Castle WE. 1919. Is the arrangement of the genes in the chromosome linear? *Proc Natl Acad Sci U S A*. 5(2):25–32.

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

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450(7167):203–218.

Cleveland WS. 1979. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc*. 74(368):829–836.

Colombo PC, Jones GH. 1997. Chiasma interference is blind to centromeres. *Heredity* 79(2):214–227.

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

Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet*. 14(4):262–274.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158.

Dumont BL, Broman KW, Payseur BA. 2009. Variation in genomic recombination rates among heterogeneous stock mice. *Genetics* 182(4):1345–1349.

Dumont BL, White MA, Steffy B, Wiltshire T, Payseur BA. 2011. Extensive recombination rate variation in the house mouse species complex inferred from genetic linkage maps. *Genome Res*. 21(1):114–125.

Felsenstein J. 1979. A mathematically tractable family of genetic mapping functions with different amounts of interference. *Genetics* 91(4):769–775.

Fishman L, Willis JH. 2005. A novel meiotic drive locus almost completely distorts segregation in *Mimulus* (monkeyflower) hybrids. *Genetics* 169(1):347–353.

Fiston-Lavier A-S, Singh ND, Lipatov M, Petrov DA. 2010. *Drosophila melanogaster* recombination rate calculator. *Gene* 463(1–2):18–20.

Gilliland WD. 2015. A comment on fine-scale heterogeneity in crossover rate in the garnet-scalloped region of the *Drosophila melanogaster* X chromosome. *Genetics* 201(3):1275–1277.

Greenwald I. 2016. WormBook: WormBiology for the 21st century. *Genetics* 202(3):883–884.

Haldane JBS. 1919. The combination of linkage values, and the calculation of distances between the loci of linked factors. *Genetics* 8(4):291–309.

Hinch AG, Zhang G, Becker PW, Moralli D, Hinch R, Davies B, Bowden R, Donnelly P. 2019. Factors influencing meiotic recombination revealed by whole-genome sequencing of single sperm. *Science* 363(6433):eaau8861.

Huehn M. 2011. On the bias of recombination fractions, Kosambi's and Haldane's distances based on frequencies of gametes. *Genome* 54(3):196–201.

Hughes SE, Miller DE, Miller AL, Hawley RS. 2018. Female meiosis: synapsis, recombination, and segregation in *Drosophila melanogaster*. *Genetics* 208(3):875–908.

Hultén MA. 2011. On the origin of crossover interference: a chromosome oscillatory movement (COM) model. *Mol Cytogenet*. 4(1):10.

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

Hunter N. 2015. Meiotic recombination: the essence of heredity. Cold Spring Harb Perspect Biol. 7(12):a016618

Ismail NIB, Kato Y, Matsuura T, Watanabe H. 2018. Generation of white-eyed *Daphnia magna* mutants lacking scarlet function. *PLoS One* 13(11):e0205609.

Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen C-F, Thomas MA, Haussler D, Jacob HJ. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res*. 14(4):528–538.

Kaur T, Rockman MV. 2014. Crossover heterogeneity in the absence of hotspots in *Caenorhabditis elegans*. *Genetics* 196(1):137–148.

Kofler R, Pandey RV, Schlötterer C. 2011. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* 27(24):3435–3436.

Kosambi DD. 1943. The estimation of map distances from recombination values. *Ann Eugen*. 12(1):172–175.

Kulathinal RJ, Bennett SM, Fitzpatrick CL, Noor MAF. 2008. Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. *Proc Natl Acad Sci U S A*. 105(29):10051–10056.

Langley CH, Lazzaro BP, Phillips W, Heikkinen E, Braverman JM. 2000. Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the *Drosophila melanogaster* X chromosome. *Genetics* 156(4):1837–1852.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 15(6):R84.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21):2987–2993.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14):1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

Martin SH, Jiggins CD. 2017. Interpreting the genomic landscape of introgression. *Curr Opin Genet Dev*. 47:69–74.

Mather K. 1939. Crossing over and heterochromatin in the X chromosome of *Drosophila melanogaster*. *Genetics* 24(3):413–435.

McCune AR, Fuller RC, Aquilina AA, Dawley RM, Fadool JM, Houle D, Travis J, Kondrashov AS. 2002. A low genomic number of recessive lethals in natural populations of bluefin killifish and zebrafish. *Science* 296(5577):2398–2401.

McDonald IC, Overland DE, Leopold RA, Degrugillier ME, Morgan PB, Hofmann HC. 1975. Genetics of house flies. Variability studies with North Dakota, Texas, and Florida populations. *J Hered*. 66(3):137–140.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20(9):1297–1303.

McPeek MS, Speed TP. 1995. Modeling interference in genetic recombination. *Genetics* 139(2):1031–1044.

McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* 304(5670):581–584.

Meisel R P, Gonzales C A, Luu H. 2017. The house fly Y Chromosome is young and minimally differentiated from its ancient X Chromosome partner. *Genome Res*. 27(8):1417–1426.

Miller DE, Staber C, Zeitlinger J, Hawley RS. 2018. Highly contiguous genome assemblies of 15 species generated using nanopore sequencing. *G3 (Bethesda)* 8(10):3131–3141.

Miller DE, Takeo S, Nandanan K, Paulson A, Gogol MM, Noll AC, Perera AG, Walton KN, Gilliland WD, Li H, et al. 2012. A whole-chromosome analysis of meiotic recombination in *Drosophila melanogaster*. *G3 (Bethesda)* 2(2):249–260.

Morgan TH. 1911a. The application of the conception of pure lines to sex-limited inheritance and to sexual dimorphism. *Am Nat*. 45(530):65–78.

Morgan TH. 1911b. Random segregation versus coupling in Mendelian inheritance. *Science* 34(873):384–384.

Muller HJ. 1916. The mechanism of Crossing-Over. *Am Nat*. 50(592):193–221.

Nachman MW. 2002. Variation in recombination rate across the genome: evidence and implications. *Curr Opin Genet Dev*. 12(6):657–663.

Neel JV. 1941. A relation between larval nutrition and the frequency of crossing over in the third chromosome of *Drosophila melanogaster*. *Genetics* 26(5):506–516.

Page SL, Hawley RS. 2004. The genetics and molecular biology of the synaptonemal complex. *Annu Rev Cell Dev Biol*. 20(1):525–558.

Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. 2019. A review of spline function procedures in R. *BMC Med Res Methodol*. 19:46.

Ramos AP, Gustafsson O, Labert N, Salecker I, Nilsson D-E, Averof M. 2019. Analysis of the genetically tractable crustacean *Parhyale hawaiensis* reveals the organisation of a sensory system for low-resolution vision. *BMC Biol*. 17(1):67.

Redfield H. 1966. Delayed mating and the relationship of recombination to maternal age in *Drosophila melanogaster*. *Genetics* 53(3):593–607.

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

Rockman MV, Kruglyak L. 2009. Recombinational landscape and population genomics of *Caenorhabditis elegans*. *PLoS Genet*. 5(3):e1000419.

Russell JJ, Theriot JA, Sood P, Marshall WF, Landweber LF, Fritz-Laylin L, Polka JK, Oliferenko S, Gerbich T, Gladfelter A, et al. 2017. Non-model model organisms. *BMC Biol*. 15(1):55.

Schulte C, Theilenberg E, Müller-Borg M, Gempe T, Beye M. 2014. Highly efficient integration and expression of piggyBac-derived cassettes in the honeybee (*Apis mellifera*). *Proc Natl Acad Sci U S A*. 111(24):9003–9008.

Singh ND, Criscoe DR, Skolfield S, Kohl KP, Keebaugh ES, Schlenke TA. 2015. Fruit flies diversify their offspring in response to parasite infection. *Science* 349(6249):747–750.

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

Singh SR, Coppola V, editors. 2014. Mouse genetics: methods and protocols. New York: Humana Press.

Smukowski CS, Noor MAF. 2011. Recombination rate variation in closely related species. *Heredity* 107(6):496–508.

Soetaert K, Herman PMJ, editors. 2009. A practical guide to ecological modelling: using R as a simulation platform. Dordrecht (the Netherlands): Springer.

Stephan W. 2019. Selective sweeps. *Genetics* 211(1):5–13.

Stern C. 1926. An effect of temperature and age on Crossing-Over in the first chromosome of *Drosophila melanogaster*. *Proc Natl Acad Sci U S A*. 12(8):530–532.

Stevison LS, Noor MAF. 2010. Genetic and evolutionary correlates of fine-scale recombination rate variation in *Drosophila persimilis*. *J Mol Evol*. 71(5–6):332–345.

Sturtevant AH. 1913. The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *J Exp Zool*. 14(1):43–59.

Sturtevant AH, Bridges CB, Morgan TH. 1919. The spatial relations of genes. *Proc Natl Acad Sci U S A*. 5(5):168–173.

Sun H, Rowan BA, Flood PJ, Brandt R, Fuss J, Hancock AM, Michelmore RW, Huettel B, Schneeberger K. 2019. Linked-read sequencing of gametes allows efficient genome-wide analysis of meiotic recombination. *Nat Commun*. 10(1):4310.

Tan Y-D, Fornage M. 2008. Mapping functions. *Genetica* 133(3):235–246.

Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, Matthews BB, Millburn G, Antonazzo G, Trovisco V, et al. 2019. FlyBase 2.0: the next generation. *Nucleic Acids Res*. 47(D1):D759–D765.

Tilk S, Bergland A, Goodman A, Schmidt P, Petrov D, Greenblum S. 2019. Accurate allele frequencies from ultra-low coverage Pool-Seq samples in Evolve-and-Resequence experiments. *G3 (Bethesda)* 9(12):4159–4168.

Umehara T, Tsujita N, Shimada M. 2019. Activation of toll-like receptor 7/8 encoded by the X chromosome alters sperm motility and provides a novel simple technology for sexing sperm. *PLoS Biol*. 17(8):e3000398.

Wang Y, Rannala B. 2008. Bayesian inference of fine-scale recombination rates using population genomic data. *Philos Trans R Soc Lond B Biol Sci*. 363(1512):3921–3930.

Wei KH-C, Reddy HM, Rathnam C, Lee J, Lin D, Ji S, Mason JM, Clark AG, Barbash DA. 2017. A pooled sequencing approach identifies a candidate meiotic driver in *Drosophila*. *Genetics* 206(1):451–465.

Yasukochi Y. 1998. A dense genetic map of the silkworm, *Bombyx mori*, covering all chromosomes based on 1018 molecular markers. *Genetics* 150(4):1513–1525.