Predictive Models of Recombination Rate Variation across the *Drosophila melanogaster* Genome

Andrew B. Adrian[1,†], Johnny Cruz Corchado[2,†], and Josep M. Comeron[1,2,*]

[1]Department of Biology, University of Iowa

[2]Interdisciplinary Graduate Program in Genetics, University of Iowa

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: josep-comeron@uiowa.edu.

Accepted: July 25, 2016

Abstract

In all eukaryotic species examined, meiotic recombination, and crossovers in particular, occur non-randomly along chromosomes. The cause for this non-random distribution remains poorly understood but some specific DNA sequence motifs have been shown to be enriched near crossover hotspots in a number of species. We present analyses using machine learning algorithms to investigate whether DNA motif distribution across the genome can be used to predict crossover variation in *Drosophila melanogaster*, a species without hotspots. Our study exposes a combinatorial non-linear influence of motif presence able to account for a significant fraction of the genome-wide variation in crossover rates at all genomic scales investigated, from 20% at 5-kb to almost 70% at 2,500-kb scale. The models are particularly predictive for regions with the highest and lowest crossover rates and remain highly informative after removing sub-telomeric and -centromeric regions known to have strongly reduced crossover rates. Transcriptional activity during early meiosis and differences in motif use between autosomes and the *X* chromosome add to the predictive power of the models. Moreover, we show that population-specific differences in crossover rates can be partly explained by differences in motif presence. Our results suggest that crossover distribution in *Drosophila* is influenced by both meiosis-specific chromatin dynamics and very local constitutive open chromatin associated with DNA motifs that prevent nucleosome stabilization. These findings provide new information on the genetic factors influencing variation in recombination rates and a baseline to study epigenetic mechanisms responsible for plastic recombination as response to different biotic and abiotic conditions and stresses.

Key words: Key Words: recombination, double strand break, DNA motif analysis, machine-learning algorithms.

Introduction

Meiosis is a pervasive process among eukaryotes and the meiotic machinery is heavily conserved (Keeney 2001). Yet the rate of meiotic recombination, and crossover in particular, exhibits an astounding degree of variation across genomes as well as between closely related species, populations of the same species, and even among individuals of the same population (Neel 1941; Parsons 1988; Kim et al. 2007; Coop et al. 2008; Kulathinal et al. 2008; Mancera et al. 2008; Kong et al. 2010; Dumont et al. 2011; Fledel-Alon et al. 2011; Ross et al. 2011; Smukowski and Noor 2011; Comeron et al. 2012; McGaugh et al. 2012; Miller et al. 2012; Singh et al. 2013; Gossmann et al. 2014; Liu et al. 2015). Moreover, both the overall number of crossovers and their distribution across genomes are affected by other factors such as age, temperature, food, and stressors, indicating that a precise description of crossover distribution requires characterizing genetic and epigenetic factors (Stern 1926; Neel 1941; Redfield 1966; Brooks 1988; Parsons 1988; Kong et al. 2002; Hussin et al. 2011; Singh et al. 2015).

To gain insight into the factors involved in crossover localization much attention has been given to short DNA sequence motifs near crossovers. Computational analyses of high-resolution crossover maps can identify specific motifs enriched at hotspot regions, but analyses of motif presence are rarely *predictive* enough to forecast patterns of crossover variation at a whole-genome scale. One of these cases is the 13-mer DNA motif recognized by the histone methyltransferase PRDM9 in humans and mice, with PRDM9 promoting histone methylation and meiotic crossover around the motif (Baudat et al. 2010; Parvanov et al. 2010; Billings et al. 2013). The PRDM9-associated motif is very highly significantly enriched near human crossovers, being present in approximately 40–60% of crossover hotspots (Myers et al. 2008; Hinch, et al. 2011). The reverse is much less often true: the presence of the PRDM9-associated motif is not a strong predictor of crossover distribution across the genome (Ségurel et al. 2011). A recent analysis across the ape phylogeny supports this conclusion, with enrichment of putative PRDM9 binding in recombination hotspot regions but no association between PRDM9 presence and local increases in recombination rates when measured broadly across the genome (Stevison et al. 2015). An equivalent case is observed in the yeast *Schizosaccharomyces pombe* where motifs enriched near some hotspots are, nonetheless, very poor predictors of hotspot localization genome-wide (Fowler et al. 2014).

In *Drosophila*, high-resolution recombination maps have revealed that crossover rates can vary 20- to 40-fold across genomic regions traditionally assumed to exhibit limited variation in recombination rates (Kulathinal et al. 2008; Comeron et al. 2012; McGaugh et al. 2012; Miller et al. 2012; Singh et al. 2013). These studies describe peaks of crossover rates across *Drosophila* genomes that are far less extreme and physically discrete than in species with more traditional hotspots, where crossover rates are >100-fold higher than in adjacent regions (Ségurel et al. 2011). Moreover, *Drosophila* species, like other species including some placental mammals, do not have functional PRDM9 orthologs (Oliver et al. 2009; Parvanov et al. 2010; Muñoz-Fuentes et al. 2011; Heil and Noor 2012). Predictably, the 13-mer motif associated with human hotspots that is recognized by PRDM9 is not observed near crossover events in *Drosophila* (Comeron et al. 2012; Heil and Noor 2012). In fact, sequence analyses in *Drosophila melanogaster* have identified not one but many DNA motifs significantly enriched near crossover events (Comeron et al. 2012; Miller et al. 2012; Singh et al. 2013). Combined, current data suggest that *Drosophila* has no traditional hotspots and we hypothesized that crossover-associated DNA motifs could be more evolutionary stable than in primates where hotspots become inactive relatively young (Coop and Myers 2007).

Here, we investigated whether the genomic distribution of specific DNA motifs has predictive power in describing crossover landscapes across the genome of *D. melanogaster.* The motifs under study were recently identified as motifs enriched near crossover sites through analysis of experimentally genotyped recombinant offspring (Comeron et al. 2012). We have now generated genome-wide landscapes of motif presence taking into account the probabilistic nature of motif sequences, background composition and a FDR-based genome-wide motif detection approach. We have also obtained genome-wide landscapes of crossover rates based on population patterns of linkage disequilibrium (LD) in two *D. melanogaster* natural populations (Chan et al. 2012). Importantly, these LD-based landscapes of crossover rates are completely independent from the data used to identify the DNA motifs, thus allowing for an unbiased evaluation of a potential association of motif distribution with crossover landscapes.

We show that the variable presence of multiple motifs across the genome generate non-linear quantitative models that explain a significant fraction of the genome-wide variation in crossover rates at all genomic scales investigated, from 20% at 5-kb to almost 70% at 2,500-kb scale. Also across all scales analyzed, the models are particularly accurate at detecting the genomic regions with the highest and lowest 10% crossover rates. Interestingly, these models remain highly predictive of crossover rate variation after removing sub-telomeric and -centromeric regions known to have strongly reduced crossover rates, and predict minimal levels of crossover across the dot (fourth) chromosome, which is known to be achiasmatic in *Drosophila* female meioses. Moreover, we report that the effect of motif presence on crossover rates differs between autosomal arms and the *X* chromosome, and show that transcriptional activity during early meiosis adds predictive power to the models thus explicitly including a potential mechanistic explanation to the known plasticity in recombination rates. Finally, we show that the most informative motifs predicting high crossover rates share properties associated with highly localized genomic regions depleted of nucleosomes.

Results

Genome-Wide Landscapes of DNA Motifs

The study of almost 2,000 crossover events mapped with high-resolution in *D. melanogaster* uncovered many DNA motifs enriched within the 500-bp sequence encompassing these crossover events (Comeron et al. 2012). These motifs, therefore, were localized to less than 1% of the genome. To generate landscapes of motif presence across the whole genome, we used the position probability matrix (PPM) of these motifs and took into account the numerous false positives expected in any large-scale genomic study as well as background nucleotide composition (see “Methods” section for details). Unless otherwise indicated, we focused on the 12 motifs present more than 1,000 times genome-wide once a 1% false discovery rate (FDR) correction is applied (see supplementary table S1 and fig. S1, Supplementary Material online for motif sequences). Motif presence ranges up to over 18,000 motif hits (M5) and highlights the need for caution when interpreting the biological relevance of individual motifs at specific genomic locations, even when FDR is set to 1%. At first glance, several motifs show variation in their distribution on a chromosome scale that visually follows the traditional distribution of crossover rate variation in *D. melanogaster* (fig. 1), with motif presence reduced near centromeric and, to a lesser degree, telomeric regions (Morton et al. 1976; Lindsley and Zimm 1992; Fiston-Lavier et al. 2010; Comeron et al. 2012; Miller et al. 2012).

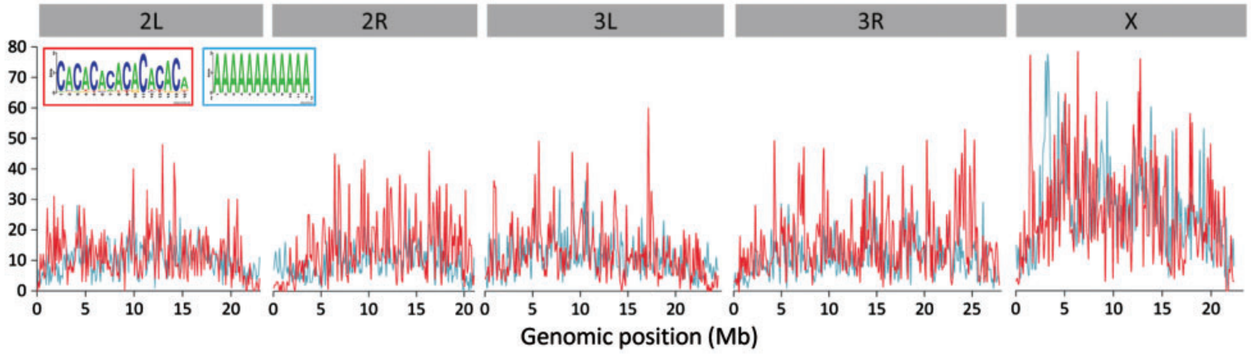

Fig. 1.—Genomic landscape of motifs. Number of motifs per 100-kb for motif 3 (M3 in blue; $[A]_n$) and motif 4 (M4 in red; $[CA]_n$) across autosomal arms 2L, 2R, 3L, 3R and the *X* chromosome (see supplementary fig. S1 for motif sequence information). Presence shown after applying a 1% FDR (see "Methods" section).

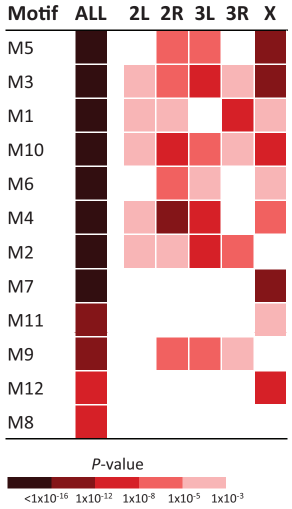

Fig. 2.—Probability heatmap of correlation between presence of individual motifs and crossover rates ρ_{LD}. *P*-values of non-parametric correlation (Spearman's r_s) of motif presence and LD-based crossover rates ρ_{LD} calculated genome-wide and for each chromosome arm. Only motifs showing correlations with $P < 0.01$ in genome-wide analyses are shown. Motifs are ordered based on Spearman's correlation across the genome, with M5 (Spearman's $r_s^2 = 0.237$, $P = 1 \times 10^{-71}$) showing the strongest correlation.

Motif Presence Is Correlated with Crossover Rates across the Genome

To study genome-wide associations between variation in motif presence and crossover rates, we obtained estimates of crossover rates that are independent of the experimental genetic map used to find the initial motifs thus preventing any circularity. We used LDhelmet (Chan et al. 2012), a method that generates high-resolution population estimates of crossover rate (ρ_{LD}) across the genome based on patterns of nucleotide polymorphism and LD. LDhelmet was applied to the *D. melanogaster* African Rwanda (RG) population [(Pool et al. 2012; Lack et al. 2015); see "Methods" setcion for details] because this population shows very low levels of admixture and is from the sub-Saharan ancestral range of *D. melanogaster*, which minimizes the non-equilibrium effects caused by recent expansion observed in western Africa and non-African populations (Pool et al. 2012; Lack et al. 2015). To investigate potential intraspecific variation in the distribution of motifs and/or crossover rates, we applied LDhelmet to the *D. melanogaster* African Zambia (ZI) population, which also shows low levels of admixture (Pool et al. 2012; Lack et al. 2015). Unless noted, analyses are reported using LD-based crossover estimates ρ_{LD} and motif distribution for the RG population at 100-kb scale.

The direct comparison of ρ_{LD} and motif presence at 100-kb scale shows positive associations for all 12 motifs analyzed ($P < 1 \times 10^{-6}$), with Spearman's r_s^2 higher than 0.10 ($P < 5 \times 10^{-31}$) for eight motifs (fig. 2 and supplementary fig. S1, Supplementary Material online). We also observe that some motifs show clear differences among chromosome arms in terms of association with ρ_{LD} (fig. 2 and supplementary fig. S2, Supplementary Material online). Variation in M7 presence shows a very strong association with ρ_{LD} along the *X* chromosome (Spearman's $r_s^2 = 0.24$, $P = 4.1 \times 10^{-15}$) whereas it shows no association along autosomal arms ($P > 0.1$ in all autosomal arms). A similar tendency can be seen for motifs M11 and M12 that also contain $[TA]_n$ repeats. M2, on the other hand, shows a positive association with ρ_{LD} along all four autosomal arms but not in the *X* chromosome. These differences are unlikely to be caused by differences in statistical power alone, as M2 is more frequently observed on the *X* chromosome than on autosomes (average presence per 100 kb is 16.4 and 8.5 for the *X* and autosomes, respectively). Although some motifs (e.g. M1, M4, M5, and M6) show diverse degrees of association with ρ_{LD} among autosomal arms, the difference in potential effects of motif presence on crossover localization seems to be mostly influenced by a dichotomy between autosomal arms and the *X* chromosome (see below).

To obtain an initial model that considers multiple motifs to explain crossover distribution, we first performed Least Absolute Shrinkage and Selection Operator (LASSO) regression (see "Methods" section for details) (Tibshirani 1996; Hastie et al. 2009). LASSO regression is a technique that favors solutions with fewer parameter values under a *linear* model, simultaneously performing variable selection and simplifying model interpretation. LASSO exposes six heavily weighted motifs (in order of importance M3, M1, M7, M5, M6, and M2; supplementary fig. S3, Supplementary Material online), all positively associated with crossover rates ρ_{LD}. With these six motifs, LASSO fits a model of motif presence that explains more than 20% of the variation in ρ_{LD} genome-wide ($R^2_{LASSO} = 0.234$, $P < 2.2 \times 10^{-16}$). Note, however, that non-parametric Spearman's correlations between motif presence and ρ_{LD} suggest that some individual motifs (e.g. M3 and M5) explain crossover variation when analyzed independently almost as much as the more complex linear LASSO model that includes several motifs (fig. 2 and supplementary fig. S1, Supplementary Material online). Easing the constraints of LASSO ($\lambda + 1$ S.E; see "Methods" section) allows all 12 motifs under study to enter the model but this highly complex and likely over-parameterized model exhibits little improvement in overall performance ($R^2_{LASSO} = 0.252$, $P < 2.2 \times 10^{-16}$).

Predictive Models of Variation in Crossover Rates across the Genome Based on Motif Presence

We applied two machine learning methods to investigate models of crossover rate variation using motif distribution as continuous predictors. We first used Random Forests (RF) (Breiman 2001; Lee et al. 2005; Banfield et al. 2007) as a form of supervised learning to discriminate between genomic regions (classes) with different crossover rates, particularly between low and high rates. We later constructed a quantitative predictive model using Multivariate Adaptive Regression Splines (MARS) (Friedman 1991; Friedman and Roosen 1995; Hastie et al. 2009) (see "Methods" section for details). MARS is an approach that allows multiple interactions among motifs and non-linear effects thus potentially capturing saturation/insensitivity and cooperative action of different motifs when describing crossover rates. Importantly, MARS allows obtaining a final explicit and continuous model of crossover rates based on the combined presence of multiple motifs.

Random Forests (RF) Categorical Modeling

We split all 100-kb regions into ten approximately equally sized bins, from the lowest (class A) to highest (class J) 10% ρ_{LD}, and applied random forests (RF) to classify crossover classes using a 10-fold cross-validation approach (see "Methods" section). The correctness of the RF models is measured by accuracy (true positive rate) and the area under the curve (AUC) that indicates the ability of the model to discriminate between the different classes, with AUC scores ranging from 0.5 (indicating that a model has no discriminatory ability) to 1 (indicating that the model can discriminate perfectly among all classes). Note that RF does not directly generate probability values associated with the whole model. We, therefore, obtained the statistical significance of RF models by comparing the accuracy and AUC generated by the model and the accuracy and AUC generated by RF when estimates of ρ_{LD} are randomized among 100-kb regions (250,000 randomizations per model).

Using the ten ρ_{LD} classes as our class variables, we applied RF with motif distribution across the genome as continuous predictors to later add transcription data and chromosome arm. Genome-wide, motif distribution allows RF to generate a model with an accuracy of 23.1% (vs. a random expectation of 10%, $P < 3 \times 10^{-6}$) and a mean AUC = 0.644 ($P < 3 \times 10^{-6}$). Although this RF model includes all 12 motifs with significant effects on the model, the six more important motifs based on information gain (M5, M4, M3, M1, M2, and M7) can explain more than 85% of the model. Overall, the model performs fundamentally well for the top and bottom classes (fig. 3A), with accuracy of 78.0 and 56.3%, respectively (more than 5-fold enriched, $P < 3 \times 10^{-6}$, in both cases). Enrichment based on AUC shows an equivalent pattern, with AUC of 0.876 and 0.814 for the top and bottom 10% classes, respectively ($P < 3 \times 10^{-6}$ in both cases). This study of motif presence correctly classifies ρ_{LD} class within one step of their true class 44% of the time, indicating that when our model fails to accurately predict a class, it often falls into the adjacent bin.

A RF model with motif distribution and information on transcribed genes during early meiosis (Adrian and Comeron 2013) increases the genome-wide accuracy to 24.4% and AUC to 0.678 ($P < 3 \times 10^{-6}$ in both cases), whereas motifs, transcription data and chromosome arm information generate an improved model with accuracy and AUC of 27% and 0.727, respectively ($P < 3 \times 10^{-6}$ in both cases). In contrast, genomic properties such as the number of annotated genes or proportion of exonic sequence in each window do not significantly affect model accuracy (data not shown) and these variables are never within the top ten variables in term of importance within the model. Similarly, GC content per window is neither ranked highly by information gain criteria nor does it have a substantial impact on classification accuracy.

The application of RF to each chromosomal arm separately reveals that the models are much more accurate predicting ρ_{LD} variation for the *X* chromosome (51.8% accuracy, a 5-fold enrichment relative to random expectations) than for autosomal arms (21.0% with accuracy, ranging between 16.5% for 2L and 24.5% for 3L) (fig. 3B). An equivalent conclusion is obtained based on AUC, with AUC of 0.68 for the *X*

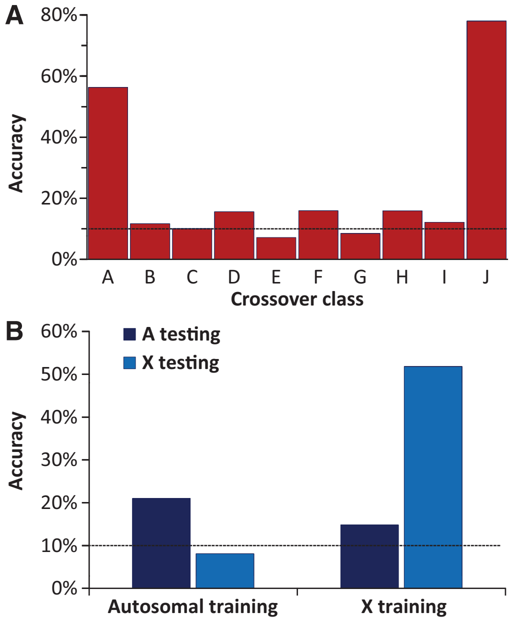

Fig. 3.—Random Forest (RF) models. A) Accuracy (true positive rate) is given for 10 crossover classes, from class A (regions with lowest 10% crossover rate ρ_{LD}) to class J (regions with highest 10% crossover rate ρ_{LD}). Random accuracy (uninformative model) per class is 10% (horizontal dashed line). The model tested utilizes all 12 motifs to predict crossover classes (see "Methods" section for details). B) Accuracy when the model is trained with data from one autosomal arm and applied to either other autosomal arms or to the *X* chromosome (left) as testing set, or trained with data from the X chromosome and applied to either autosomal arms or to the *X* chromosome (right).

chromosome whereas AUC ranges between 0.51 (2L and 2R) and 0.62 (3L) for autosomes. In agreement with our initial analyses, RF also indicates that the influence of some motifs on crossover distribution is not equivalent genome-wide. M5 and M7 are more important motifs in the RF model for the *X* chromosome (combining 33% of the relative importance of all motifs) than in the model for autosomes (combining 18% of the relative importance). Conversely, M4 and M2 are ranked lower in models for the *X* chromosome than in models describing crossover across autosomes (14 and 28% of the relative importance of all motifs, respectively). To further test the difference between the *X* chromosome and autosomes we applied RF using data from individual chromosomal arms as a training set and tested the accuracy of the model describing crossover rates along *other* chromosomal arms (see "Methods" section). The results of this analysis evidence that autosomal arms are worse at predicting ρ_{LD} along the *X* chromosome than along other autosomal arms (fig. 3B). Equivalently, the use of *X* chromosome data as training set generates models of ρ_{LD} variation along autosomes with an average accuracy of only 14.8%.

MARS Modeling

We used the 10-fold cross-validation approach (see "Methods" section for details) and evaluated the quality of MARS models focusing on R^2_{CV} scores (the MARS estimate of how well this model would perform on new data when using the 10-fold cross-validation approach; see "Methods" section). The simplest model that considers only the variable presence of motifs across the genome is able to explain ~50% ($R^2_{CV} = 0.501$) of the genome-wide variation in ρ_{LD} (fig. 4 and supplementary table S2, Supplementary Material online). This predictive model improves even further when including information on transcription ($R^2_{CV} = 0.512$) or chromosome arm ($R^2_{CV} = 0.540$).

The most complete MARS model, which includes motif presence, transcription data and chromosome arms, explains 57% ($R^2_{CV} = 0.569$) of the variation in crossover genome-wide. This most complex model identifies chromosome arm, number of actively transcribed genes and seven motifs as significantly important within the model (M5, M3, M4, M7, M2, M12, and M10), several of which exhibit non-linear effects on crossover rates (fig. 4B and C). Note that MARS estimates based on less conservative methods such as the legacy mode (R^2_{Legacy}) would imply an even higher influence of motif presence on ρ_{LD} (R^2_{Legacy} values always larger than R^2_{CV}; supplementary table S2, Supplementary Material online) but caution should be applied to the interpretation of these high estimates due to overfitting (see "Methods" section).

Putative Influence of the Method to Estimate Crossover Rates

Supplementary figure S1, Supplementary Material online shows Spearman's correlation between motif distribution and crossover rates across the genome based on either population estimates (ρ_{LD}) or from direct genotyping of offspring [i.e., experimentally identified crossovers; (Comeron et al. 2012)]. We observe that all 12 DNA motifs show significant correlations between presence across the genome and experimentally generated crossover landscapes, with the highest Spearman's r_s^2 of 0.176 for M3 ($P = 7 \times 10^{-52}$) at 100-kb scale. Additionally, MARS modeling shows an important predictive power of motifs explaining these crossover rates genome-wide (R^2_{CV} ranging between 0.263 and 0.304). These data show that the observed significant contribution of motif distribution to variation in crossover rates across the genome is not exclusive to using ρ_{LD}. However, we also detect that the predictive power is higher when using population estimates ρ_{LD} and motif distribution from the same population than when using crossover rates estimated experimentally and motifs from either RG or ZI population (or reference genome). This difference can be interpreted as consequence of the admixed source of strains used to generate a recombination map for the *D. melanogaster* species and/or the environmental

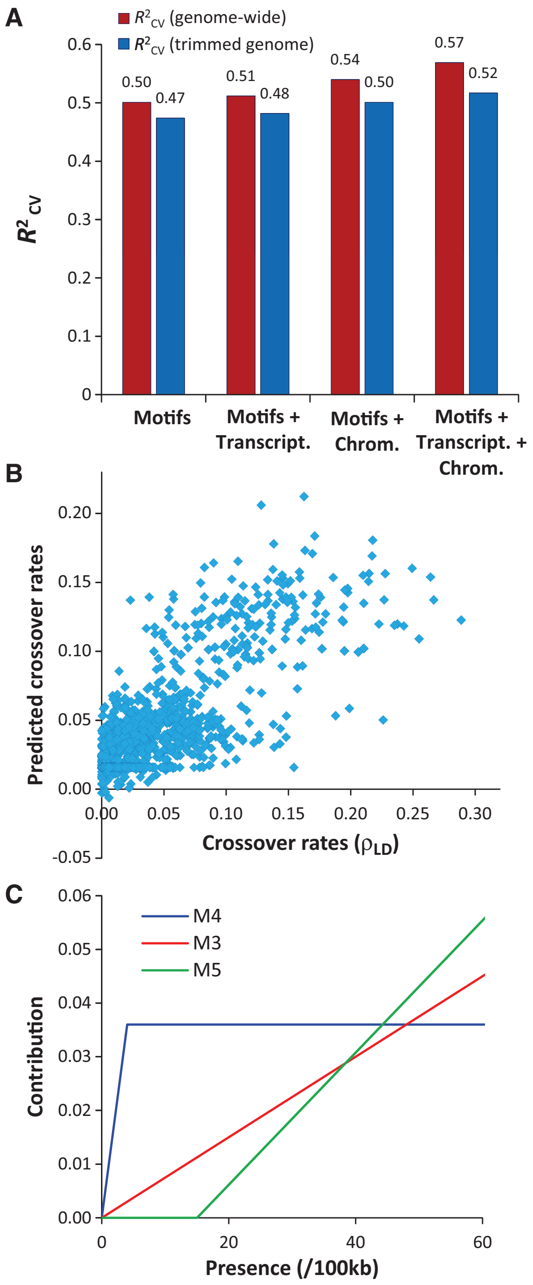

Fig. 4.—MARS predictive models of crossover rates. (A) Estimates of the predictive power of MARS models (R^2_{GCV}) based on the presence of motifs across the genome, transcription data during early meiosis, and/or chromosome arms as predictive variables. Results shown for genome-wide analyses (red) and for trimmed (after removing sub-telomeric and -centromeric regions) genome (blue). (B) Relationship between crossover rates obtained from population genetic analyses of linkage disequilibrium (*x*-axis) and those predicted based on a MARS model (*y*-axis) including motif, transcription and chromosome arm data. The unit of crossover rate is ρ_{LD} ($\rho_{LD} = 2\ N_e\ r$), where N_e is the effective population size and *r* is the rate of crossover per bp and generation in females (Chan et al. 2012). C) Examples of the linear and non-linear influence of motif presence on crossover rates. All results are shown for analyses of the RG population at 100-kb scale.

conditions when obtaining recombinants and associated epigenetic factors influencing crossover distribution.

Alternatively, one could argue that the difference between experimental and LD-based crossover landscapes exposes biases in LDhelmet when generating ρ_{LD}. In *Drosophila*, where males do not recombine, ρ_{LD} is formally equal to 2 N_e *r*, where N_e is the effective population size and *r* is the rate of crossover per bp and generation in females (see "Methods" section for details). Variation in ρ_{LD} could then reflect differences in $N_{e,}$ additional to (or rather than) differences in *r*, due to demographic events or the interplay between selection and linkage (Wright 1938; Hill and Robertson 1966; Smith and Haigh 1974; Slatkin and Hudson 1991; Charlesworth et al. 1993; Gillespie 2000; Sella et al. 2009; Cutter and Payseur 2013; Charlesworth and Campos 2014; Comeron 2014). Chan et al. (2012) showed that LDhelmet generates estimates of ρ_{LD} that are good estimators of *r* under several demographic and selective conditions, although complex adaptive selective scenarios cause more noise in the estimates of ρ_{LD} and could, potentially, affect ρ_{LD} differentially across the genome. To investigate this possibility, we focused on the ratio of ρ_{LD} (2 N_e *r*) to the amount of silent nucleotide polymorphism (π_{sil}), which is also proportional to N_e ($\pi_{sil} = 4N_e\mu$ for autosomes and $3N_e\mu$ for the *X*, where μ is the mutation rate per bp and generation). We show that the correlation between motif presence and ρ_{LD}/π_{sil} (which is independent of N_e) is equivalent to that observed when using ρ_{LD}, and this is the case at different genomic scales (supplementary table S3, Supplementary Material online). RF based on motif distribution generates models with 22.9% accuracy (AUC of 0.638) when assigning regions to their corresponding decile ρ_{LD}/π_{sil} class, with accuracy for the top and bottom 10% classes of 83.2 and 55.5%, respectively, thus with all results virtually equivalent to those obtained when using ρ_{LD}. Also equivalent is the increased accuracy (25.1%) and AUC (0.707) when including chromosome arm as variable. Furthermore, MARS generates models that explain as much of the variance in ρ_{LD}/π_{sil} across the genome as the models for ρ_{LD}, with R^2_{CV} ranging between 0.513 (models using only motifs distribution) and 0.583 (when also including transcription and chromosome data) for analyses at 100-kb scale. Combined, these data indicate that our results and conclusions supporting a strong role of motifs in crossover distribution are unlikely to be caused by the use of ρ_{LD}.

The Centromere-Effect

In *D. melanogaster*, crossover frequency is severely reduced in euchromatic regions near centromeres and, to a lesser degree, near telomeres [the so-called "centromere effect" (Beadle 1932)]. Crossover in these large sub-centromeric and telomeric regions is likely influenced by mechanisms at least partially different to those acting along the central regions of

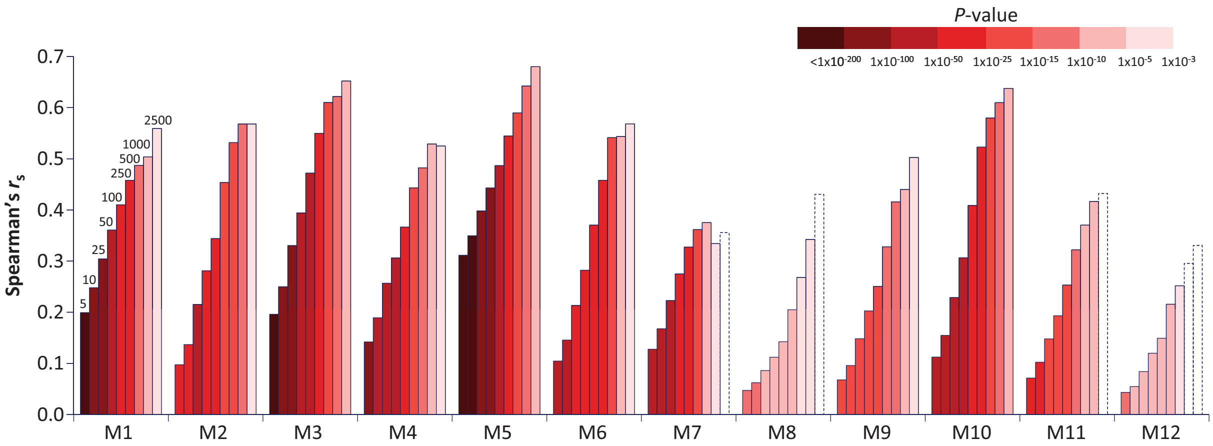

Fig. 5.—Influence of genomic scale on the correlation between motif presence and crossover rates ρ_{LD}. Spearman's non-parametric correlation (r_s) between motif presence and LD-based crossover rates ρ_{LD} is shown for intervals of 5-, 10-, 25-, 50-, 100-, 250-, 500-, 1,000-, and 2,500-kb. For each motif, the nine adjacent vertical bars indicate the different scales, from the finest scale (5-kb; left) to the broadest scale (2,500-kb; right). The color of each bar indicates the probability (*P*) of the correlation, with more significant correlations (lower probabilities) in darker red. Vertical bars with dashed borders indicate correlations with probabilities greater than 0.001. See supplementary figure S1 for motif sequence information.

chromosomal arms. In agreement, RF assigns all 100-kb regions correctly to their corresponding decile class (the majority of which are class A or B), and MARS models generate predictions of ρ_{LD} for these sub-centromeric and telomeric regions that are significantly lower than the rest of the euchromatic genome (Mann–Whitney U Test, $P < 1 \times 10^{-16}$). We, then, investigated whether motif distribution is also informative predicting variation in ρ_{LD} after removing sub-centromeric and -telomeric regions (i.e., trimmed genome) from the study (fig. 4A; see "Methods" section for details). RF across the trimmed genome shows 26.0% accuracy and 0.656 AUC. MARS generates models with R^2_{CV} of 0.482 (only motif presence) and 0.517 (motifs + chromosome arms + transcription activity during early meiosis), respectively. These results evidence that variation in motif distribution maintains significant power explaining variation in crossover rates beyond the centromere effect.

The Achiasmate Fourth Chromosome

The fourth (or dot) chromosome in *D. melanogaster* represents approximately 3.5% of the genome, consists of interspersed heterochromatic and euchromatic DNA (Sun et al. 2000), and does not experience crossing over (Hochman 1973). We asked whether our models based on DNA motif presence predict low or no crossover along this chromosome by applying RF and MARS models that were trained on genome data that did not contain the fourth chromosome and then tested these models with the fourth chromosome. RF correctly classifies all thirteen 100-kb intervals of the fourth chromosome as belonging to class A, the lowest 10% ρ_{LD} genome-wide. The application of MARS modeling using motif presence generates predictions of ρ_{LD} for these thirteen 100-kb intervals that are significantly lower than the rest of the genome (Mann–Whitney U Test, $P = 0.0022$).

Influence of Genomic Scale on the Association between Motif Presence and Crossover Rates

To investigate whether the influence of motif presence on ρ_{LD} is scale-specific, we generated data at nine different genomic scales, from fine (5-kb) to broad (2,500-kb) non-overlapping intervals across the whole genome. Figure 5 shows Spearman's r_s correlations and corresponding probability values for each of the 12 motifs analyzed at these different scales. At the finest scale of 5-kb, all motifs show a highly significant correlation between motif presence and ρ_{LD} (from $P = 3 \times 10^{-11}$ for M12 to $P < 1 \times 10^{-300}$ for M5), with Spearman's r_s ranging between 0.043 (M12) and 0.31 (M5). Motifs show increasing r_s with successively broader scales (with the only exceptions of M4 and M7 at or above 1,000-kb). Note, however, that there is a reduction in statistical significance with broader scales, at least in part expected due to the smaller number of intervals investigated, and four motifs (M7, M8, M11, and M12) show non-significant correlations with crossover at 2,500-kb scale. The application of RF and MARS at these different genomic scales shows equivalent patterns (fig. 6). There is a tendency for increased predictive power (accuracy in RF and R^2_{CV} in MARS models) with successively broader scales between 5- and 1,000-kb, whereas RF shows a clear reduction in accuracy above the 1,000-kb scale evidencing the influence of larger chromosomal properties.

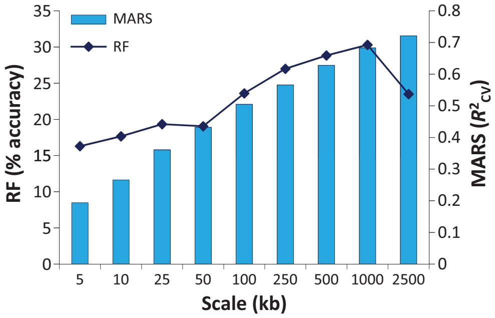

Fig. 6.—Influence of genomic scale on RF and MARS analyses. Predictive power (accuracy and R^2_{CV} in RF and MARS models, respectively) for models using motif presence to predict crossover rates ρ_{LD} at nine different genomic scales across the genome.

Intra-Specific Variation in Motif Presence and Crossover Rates

In many species, including *D. melanogaster*, crossover rates are known to vary among individuals of the same species in terms of total rates as well as for crossover distribution along chromosomes (Brooks and Marks 1986; Neumann and Jeffreys 2006; Graffelman et al. 2007; Coop et al. 2008; Dumont et al. 2009; Kong et al. 2010; Comeron et al. 2012; Miller et al. 2012; Bauer et al. 2013; Hunter et al. 2016). All our previous analyses focused on motif presence calculated across the genomes of individuals of the African *D. melanogaster* RG population and estimates of crossover rates ρ_{LD} for this same population. Analyses based on motif presence and estimates of ρ_{LD} using the Zambia (ZI) population generate similar conclusions (supplementary fig. S1 and Table S2, Supplementary Material online), with variation in motif presence across ZI genomes being able to explain a large fraction of the observed intra-genomic variation in ρ_{LD}. MARS, for instance, generates models of ρ_{LD} variation when studying the ZI population with higher R^2_{CV} than those obtained from the study of the RG population ($R^2_{CV} > 0.59$ for analyses at 100-kb scale, supplementary table S2, Supplementary Material online).

The comparison of the RG and ZI populations reveals differences in ρ_{LD} landscapes (Spearman's $r_s^2 = 0.52$ at 100-kb scale) as well as differences in motif distribution (Spearman's r_s^2 ranges between 0.581 for M12 and 0.979 for M4 at 100-kb scale). To investigate whether inter-population variation in motif presence plays a role in the difference in crossover landscapes, we applied MARS using different source of motif data and at different genomic scales (fig. 7). MARS models generate greater R^2_{CV} predicting ρ_{LD} variation in the RG population when using motif distribution from this same RG population than when using motif distribution across genomes of the ZI population at all scales, most noticeably at 100-kb or finer scales. The use of motif distribution of the *D. melanogaster* reference genome shows even more reduced power forecasting ρ_{LD} in the RG population, which would be consistent with its low percentage (9%) of African ancestry (Pool 2015). Combined, these analyses show that population-specific differences in crossover rates can be, to some degree, explained by population-specific differences in motif presence.

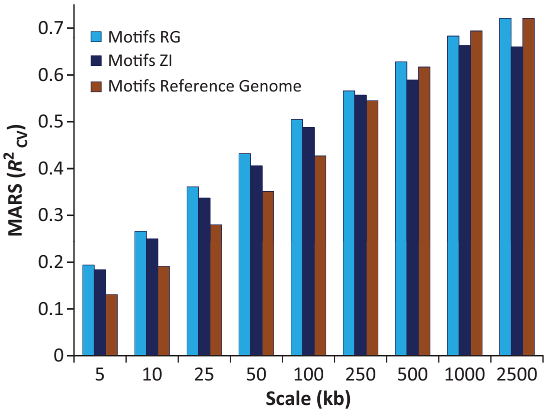

Fig. 7.—Influence of intra-specific origin of motif data and genomic scale on MARS analyses. Predictive power (R^2_{CV}) for MARS models of crossover rate ρ_{LD} across sequences of the RG population using motif distribution estimated from sequences of the same RG population, the ZI population or the *D. melanogaster* reference genome.

Predictive Motifs, Chromatin Accessibility and High-Resolution Crossover Rates

Nucleosome-depleted regions (NDR) in both *S. pombe* and *Saccharomyces cerevisiae* are strongly associated with crossovers hotspots and local patterns of double-strand break (DSB) distribution, evidencing an influence of chromatin accessibility on DSBs (Ohta et al. 1994; Wu and Lichten 1994; Hirota et al. 2007; Berchowitz et al. 2009; de Castro et al. 2012; Yamada and Ohta 2013). Importantly, however, nucleosome depletion is not informative enough to describe DSB distribution genome-wide in these yeast species (de Castro et al. 2012; Fowler et al. 2014). Most of the motifs with strong influence on crossover rates genome-wide within our RF and MARS models contain $[CA]_n$ (M1 and M4), short poly-A/T (M3 and M5) and $[AT]_n$ (M7) tracts. All these specific repeats have been proposed to prevent nucleosome stabilization, potentially increasing highly localized accessibility (Travers 1990; Perez-Martin and de Lorenzo 1997; Shimizu et al. 2000; Segal et al. 2006; Struhl and Segal 2013). We, thus, asked whether these motifs are indeed both enriched in highly accessible chromatin regions and show higher crossover rates at very local scale. We argue that, contrary to transcription-associated NDRs, accessible chromatin due to short sequences that prevent nucleosome stabilization should be constitutive and, as such, they should be a detectable component in different cells

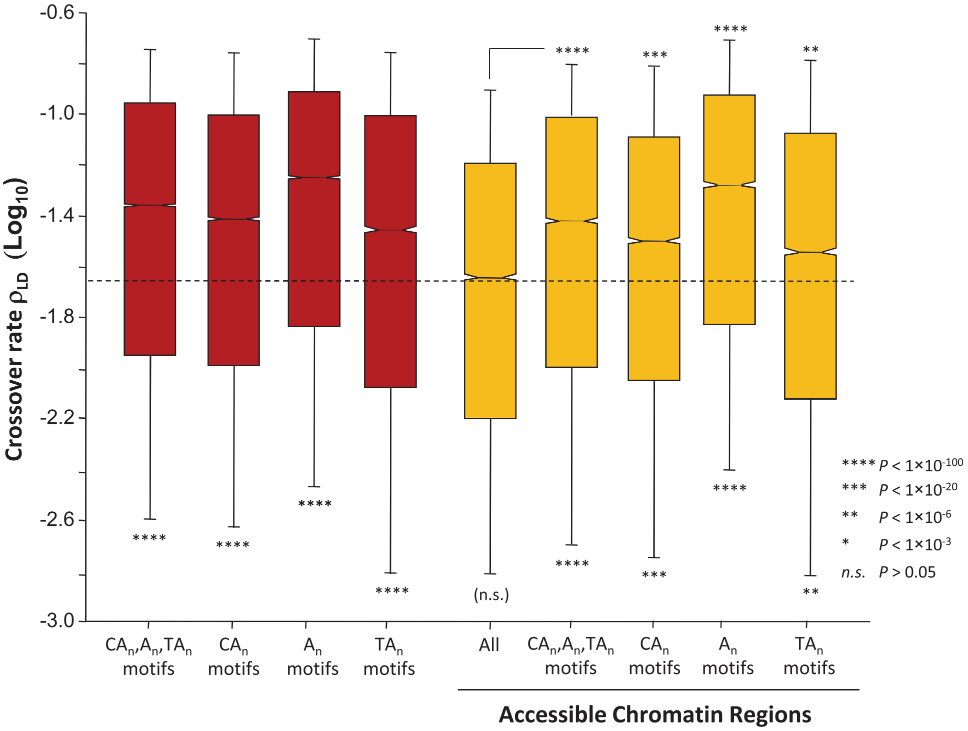

Fig. 8.—Boxplots of local crossover rate (ρ_{LD}) at motifs. Crossover rates estimated by LDhelmet (Chan et al. 2012) between SNPs surrounding the motif location in the RG population. The average distance between SNPs surrounding motifs in the RG population is 41-bp. Median crossover rates are identified by the horizontal line inside each box and the length of the box and whiskers indicate 50% and 90% CI, respectively. The horizontal dashed line indicates the genome-wide median ρ_{LD} ($\rho_{LD} = 0.0225$). Asterisks below boxes indicate the probability of having crossover rates compatible with genome-wide estimates. For analyses of local crossover rates at accessible chromatin regions, ρ_{LD} was estimated at the center of the region. The study of all accessible chromatin regions (see "Methods" section) reveals median ρ_{LD} of 0.0234 at these regions (indistinguishable from genome-wide rates, $P > 0.50$). Asterisks above boxes of accessible chromatin regions containing specific motifs ($[CA]_n$, short poly-A and/or $[TA]_n$) indicate the probability of having crossover rates compatible with all accessible chromatin regions.

types, tissues and developmental stages. Therefore, we used publicly available high-resolution open chromatin profiles generated from four *D. melanogaster* samples, including eye-antennal imaginal discs and three different cell lines (S2, BG3 and Kc167) [(Kharchenko et al. 2011; Thomas et al. 2011; Davie et al. 2015); see "Methods" section for details].

We observe that there is a significant enrichment of these five motifs in open chromatin regions compared with random expectations in each of the four samples analyzed ($P < 1 \times 10^{-53}$ in all cases). This enrichment is particularly evident for $[CA]_n$-related M1 and M4 motifs ($P < 1 \times 10^{-110}$ in all four samples) and, to a lesser degree, for the short poly-A/T M5 motif, which shows enrichment in three out of the four samples ($P = 3.4 \times 10^{-7}$ in eye-antennal imaginal discs, $P = 6.2 \times 10^{-}$ in Kc167 and $P = 0.02$ in S2). We then investigated crossover rates *at* the location of these motifs based on estimates of ρ_{LD} between informative SNPs encompassing the motif genomic position (the average distance between informative SNPs is only 43-bp thus this approach provides a very local estimates of crossover rates). Crossover rates are significantly higher at the location of these motifs than expected by random distribution (fig. 8), and this is the case whether the three types of motifs ($[CA]_n$, poly-A/T, and $[AT]_n$) are analyzed combined or separately ($P < 1 \times 10^{-190}$ in all cases). Finally, we also report that not all accessible chromatin regions in the samples investigated are associated with increased crossover rates. In fact, only accessible regions containing $[CA]_n$, poly-A/T, and/or $[AT]_n$ exhibit significantly increased crossover rates ($P < 1 \times 10^{-9}$ in all cases, fig. 8) as expected if they are associated with constitutive NDRs and thus relevant for female meiosis.

Discussion

In our study, we obtained FDR-corrected landscapes of motif presence across the genome using the likelihood of each *k*-mer sequence of nucleotides to correspond to a given position probability matrix (PPM) and taking into account genome size and nucleotide composition. The use of a FDR instead of a direct arbitrary probability is necessary to limit the extent of false positives and this parameter should be tuned appropriately for each study. We set FDR at 1% because it generates a large number of motif hits whereas still producing PPMs equivalent to those obtained under the most stringent FDR of 0.1%, thus suggesting that false positives do not seriously alter motif detection. We also confirmed that a single PPM per motif is adequate across the whole genome by comparing the top 50 and bottom 50 recombination-correlated regions by residual (which followed an approximately normal distribution). We found no significant differences between PPMs recovered from these two classes of regions indicating that a single PPM per motif can be used at genome-wide scale and that motif count, rather than motif sequence or quality, is a more important predictor of crossover in our data once FDR is applied.

Differences among Chromosome Arms

Our study has exposed that the association between motif presence and crossover rates differs depending on chromosome arm, mostly as a dichotomy between autosomal arms and the *X* chromosome. Our results obtained using either ρ_{LD} or ρ_{LD}/π_{sil} suggest that different genomic regions may utilize different combinations of DNA motifs as localizing factors for crossover formation (and potentially DSBs). Outside of differential gene expression patterns (Meisel and Connallon 2013; Llopart 2015) or average rates of crossover in females, it is not direct how the *X* chromosome and autosomes could be different in terms of motif utilization for meiotic DSB formation and resolution as crossovers. At this time, we hypothesize that such a potential mechanistic link could be associated with differences in chromatin exposure and transcription during early female meiosis (Adrian and Comeron 2013) and/or spatio-temporal compartmentalization of chromosome arms within the nucleus (Parvinen and Soderstrom 1976; Koszul et al. 2008; Shibuya et al. 2014) during early female meiosis. At a more practical level, our results also indicate that analyses of motif occurrence based on a single genomic region or chromosome may not be necessarily applicable genome-wide, thus explaining differences in motif detection among studies.

Crossover Localization across the Genome

Based on the knowledge obtained from this study and others, a general model is emerging where crossover distribution is determined by a combination of factors acting hierarchically at different physical scales, including motifs, transcription, and chromatin structures (Petes 2001; Kleckner 2006; Pan and Keeney 2007; Pan et al. 2011; Adrian and Comeron 2013; Borde and de Massy 2013; Smukowski Heil et al. 2015). In *D. melanogaster*, the centromere effect describes variation in crossover distribution at the largest scale (hundreds of kb), with a severe reduction in crossover rates at sub-centromeric/telomeric regions (Beadle 1932). We observe that regions proximal to telomeric and centromeric regions have fewer motifs positively associated with crossovers. Because crossovers near centromeres increase the probability of nondisjunction events at the second meiotic division (Koehler et al. 1996), it is tempting to speculate that natural selection may have played a role in the observed paucity of recombinogenic motifs in such genomic regions. Combined, our results indicate that the centromere-effect observed today in *D. melanogaster* may be the consequence of both direct mechanistic explanations as well as long-term evolutionary forces that have reduced the presence of crossover-associated motifs in these regions.

High-resolution crossover maps in a number of species (including *D. melanogaster, S. cerevisiae*, *S. pombe*, *Plasmodium falciparum* or *Apis mellifera*) have exposed multiple motifs significantly enriched near crossovers (Gerton et al. 2000; Cromie et al. 2007; Steiner et al. 2009; Jiang et al. 2011; Bessoltane et al. 2012; Comeron et al. 2012; Singh et al. 2013; Liu et al. 2015). Here, we have showed that variation in the presence of several motifs is in fact predictive of a significant fraction of the observed variation in crossover distribution across the genome of *D. melanogaster*, where traditional hotspots have not been detected, even after removing regions near centromeres and telomeres. Moreover, our study suggests that the difference in motif landscapes between populations may be associated with the observed difference in crossover landscapes between these populations, therefore exposing an important genetic component responsible for (or associated with) variation and evolution of recombination landscapes. Additionally, we observe that this influence of motif distribution on differences in crossover rates between populations is most evident at small scale (≤250-kb), thus potentially being an unappreciated aspect of the dichotomy between evolutionary conservation of recombination landscapes at very broad genomic scales and highly variable crossover rates at finer scales (Myers et al. 2005; Coop and Przeworski 2007; Coop et al. 2008; Smukowski and Noor 2011; Comeron et al. 2012; Smukowski Heil et al. 2015).

Individually, the motifs with the strongest effect could explain up to 24% (figs. 2 and 5 and supplementary fig. S1, Supplementary Material online) of the variance in crossover rates genome-wide when analyzing the RG population at the 100-kb scale (30% for the ZI population). Standard multiple linear and the more complex LASSO linear regression analyses generate models with little or no additional advantages when describing the observed variance in crossover rates across the genome. The application of more advanced techniques, however, exposes increased predictive power to describe crossover distribution when multiple motif variables are considered without the constraints of linear models. Our study shows that RF modeling is particularly accurate at predicting regions with the highest and lowest 10% crossover rates, and reliably predicts minimal crossover across the fourth chromosome based on motif distribution alone. The continuous models generated by MARS, on the other hand, are able to account for a large fraction of the variance in crossover rates across the *D. melanogaster* genome and identify significant interactions among motifs. We also show that variables describing transcriptional activity across the genome during early meiosis and, more unexpectedly, chromosome arm add to the predictive power of the models. Notably, the influence of motif distribution on crossover rates is detected at all genomic scales analyzed (from 5- to 2,500-kb) although it is most significant at or below the 1,000-kb scale.

Open Chromatin Sites

Several of the motifs with the strongest association with crossover rates share an enrichment of [CA]n, short poly-A/T and $[AT]_n$ tracts. Gel-mobility and X-ray crystallographic studies have shown that repeated instances of A/T tracts produce a non-canonical bend in the DNA helix axis, with the in-phase repetition of these elements contributing to larger overall bends (Travers 1990; Dlakic and Harrington 1998; Hizver et al. 2001). Such unusual DNA conformation of poly-A/T tracts and other AA/TT/AT repeats is a significant factor altering protein-DNA binding specificities, prevents nucleosome formation and stabilization and can, ultimately, increase accessibility (Travers 1990; Perez-Martin and de Lorenzo 1997; Shimizu et al. 2000; Segal et al. 2006; Struhl and Segal 2013). The signal associated with [CA]n motifs across the *D. melanogaster* genome is also interesting because these motifs are markers for Z-DNA regions which are not easily wrapped into nucleosomes (Garner and Felsenfeld 1987; Herbert and Rich 1996).

In this study, we have showed that in *D. melanogaster* [CA]n, short poly-A/T tracts and [TA]n motifs are both enriched in open chromatin regions and show higher crossover rates than expected. Our results in *Drosophila* at a genome-wide scale, therefore, parallel those in yeast hotspots, where short poly-A/T and [CA]n tracts are also significantly enriched (Treco and Arnheim 1986; Mancera et al. 2008; Fowler et al. 2014). Worth mentioning, the association between A/T rich motifs and crossovers is motif-specific and not a consequence of variation in large-scale nucleotide composition because there is only a nominal and negative effect of overall nucleotide A + T content on the distribution of crossover rates [Spearman's $r_s = -0.057$ ($P = 0.006$) at 100-kb scale (Comeron et al. 2012)]. Combined, our data strongly suggest that meiotic crossover localization in *Drosophila* is influenced by both chromatin dynamics associated with transcription and DNA motifs that cause constitutive open chromatin through unusual secondary and tertiary DNA structures at least partly independent of transcription, in qualitative agreement with data from *S. cerevisiae* (Berchowitz et al. 2009). Nevertheless, additional work is required in *Drosophila* to characterize nucleosome dynamics and/or higher-order chromatin domains specific to female early meiosis stages to fully ascertain similarities and differences between species (Ohta et al. 1994; Wu and Lichten 1994; Fan and Petes 1996; Mirouze et al. 2012; Yelina et al. 2012; Adrian and Comeron 2013; Soriano et al. 2013; Aymard et al. 2014).

Conclusions

Our results show that the presence of specific DNA motifs that prevent nucleosome stabilization add an important layer of information when considering genomic landscapes of crossover in *Drosophila*, likely through very local constitutive open chromatin. Additionally, we show that the distribution of these motifs varies between populations and likely play a role in the observed intra-specific differences in crossover landscapes. Beyond the presence of these specific motifs and the large-scale centromere-effect, we identify the additional influence of transcription activity during early meiosis that creates a mechanistic link between local recombination rates and epigenetic features in a manner similar to yeast hotspots (Petes 2001). We propose that our study provides a baseline for future analyses in *Drosophila* designed to characterize the genetic causes of intra- and interspecific variation in crossover rates and the epigenetic mechanisms responsible for the known variability in crossover rates as response to different biotic and abiotic conditions and stresses.

Methods

Motif Landscape Generation

We used the position probability matrix (PPM) of 20 DNA motifs identified as enriched near crossover events from (Comeron, et al. 2012) (see supplementary fig. S1 and table S4, Supplementary Material online). To generate motif frequency estimates across chromosomes (motif landscapes), we developed a suite of custom python scripts designed to take PPM generated by MEME (Bailey et al. 2009) as input, apply a sliding-windows approach to estimate the likelihood of each stretch of DNA of containing the motif, and finally apply a FDR-based threshold to classify a sequence as belonging to a motif or not. In more detail, and for each motif, we applied a genomic scan to assign a likelihood L to every k-mer sequence across the genome to fit the PPM (with k indicating the length of the motif). L_i at position i is estimated as the $\log_{10}$-transformed product of individual probabilities of the observed nucleotide j ($j = $ A, G, C, T) at position x in the k-mer being at the same position x in the PPM ($1 \leq x \leq k$). Nucleotide probabilities of 0 within the PPM were replaced by 1×10^{-10}. We then generated a genome-wide null distribution of L_i (RL_i) based on random shuffling of nucleotide and dinucleotide composition using an equivalent approach. Finally, we used the complete null distribution of RL_i to obtain a threshold for observed L_i that would represent a desired false discovery rate (FDR) or L_{FDR}. We call a motif to be present at position i only when $L_i > L_{FDR}$. This approach allows applying any arbitrary FDR and, importantly, takes into account the number of sites under study. We chose to use a conservative FDR of 1% as it maximizes dynamic range and allows recovering sequence motifs nearly indistinguishable from those produced using an FDR $< 1\%$, all while restricting the fraction of false positives to an acceptable threshold.

We generated motif landscapes with 1% FDR for the *D. melanogaster* reference genome 5.47 (dm3 assembly; http://flybase.org/) and for the RG and ZI populations [(Pool et al. 2012; Lack et al. 2015); see below] separately. Supplementary table S4, Supplementary Material online details the PPM for all motifs investigated in this study when applied to the RG population. The set of FDR-corrected motifs generate PPMs that are similar but not an exact match to the initial set of seed PPMs, which is not unexpected due to the limited number and genomic distribution of the original set of sequences analyzed. Some of the motifs reported to be enriched near crossover sites in Comeron et al. (2012) show very limited presence once the 1% FDR-correction is applied and, as consequence, motifs M13-M20 in supplementary table S4, Supplementary Material online, with fewer than 1,000 counts genome-wide, show no significant correlation with ρ_{LD} and were not included in the analyses. Unless noted, all 12 motifs present more than 1,000 times genome-wide (M1–M12) were used in the analyses (supplementary tables S1 and S4, Supplementary Material online), and the RG population was used to estimate motif distribution and ρ_{LD} (see below). Analyses towards investigating ρ_{LD} for the ZI population used motif landscapes along chromosomes from the ZI population.

Using positional information of each motif location, we generated sliding-window estimates of FDR-corrected motif presence for non-overlapping regions at different genomic scales across the genome, including 5-, 10-, 25-, 50- 100-, 250-, 500-, 1,000-, and 2,500-kb. Chromosome positions and gene annotations were based on the *D. melanogaster* dm3 assembly and annotation release 5.47 (http://flybase.org/). Sequence logos were generated using WebLogo 3 [http://weblogo.threeplusone.com (Crooks et al. 2004)].

LD-Based High-Resolution Crossover Maps

We calculated the population-scaled recombination rate for two African populations of *D. melanogaster* using the program LDhelmet (Chan et al. 2012). LDhelmet is a statistical method that allows estimating fine-scale recombination rates across genomes based on patterns of linkage disequilibrium, where the parameter estimated is the population-scaled crossover rate per bp and generation (ρ_{LD}); $\rho_{LD} = 2N_e r$, where N_e is the effective population size of the population and r is the rate of crossover per bp and generation in females. Note, therefore, that estimates of ρ_{LD} by LDhelmet represent historic averages of estimates of crossover for the population or species under analysis. Following Chan et al. (2012), we applied a block penalty of 50 and used the RG matrix. We applied effective mutation rates estimated from the RG and ZI sequences: 0.008 and 0.0071 for autosomes and the *X* chromosome, respectively, for the RG population, and 0.011 and 0.0087 for autosomes and the *X* chromosome, respectively, for the ZI population. The data was divided into overlapping blocks of 4400 SNPs, with 200 SNPs of overlap. For each block we ran LDhelmet for 3,000,000 iterations after 300,000 iterations of burn-in. Recombination maps for each chromosomal arm were analyzed as non-overlapping adjacent windows at nine different genomic scales (5-, 10-, 25-, 50-, 100-, 250-, 500-, 1,000, and 2,500-kb). To remove sub-centromeric and –telomeric regions with strongly reduced crossover rates, we classified a sub-centromeric region by starting at the centromere and moving into the chromosome arm until a minimum of 3 consecutive 100-kb windows showed $\rho_{LD} > 0.02$, and sub-telomeric regions were assigned in an equivalent manner [see (Comeron et al. 2012)].

We analyzed the Rwanda (RG) and Zambia (ZI) populations because both show very limited levels of admixture and allow analyzing a relatively large sample of strains with no chromosomal inversions (Pool et al. 2012; Lack et al. 2015). We obtained the genomic sequences from the Drosophila Genome Nexus [http://www.johnpool.net/genomes.html; (Lack et al. 2015)] and analyzed strains with no evidence of chromosomal inversions in any chromosomal arm. In total our analysis included 19 RG sequences (RG10, RG13N, RG15, RG19, RG22, RG24, RG28, RG2, RG32N, RG33, RG34, RG35, RG38N, RG39, RG4N, RG6N, RG7, and RG8) and 20 ZI sequences (ZI184, ZI250, ZI252, ZI271, ZI311N, ZI320, ZI324, ZI332, ZI344, ZI378, ZI386, ZI398, ZI402, ZI418N, ZI420, ZI455N, ZI457, ZI477, ZI517, and ZI85). Unless noted, our analyses use the RG population because the low levels of admixture have been well characterized (Pool et al. 2012) and we masked admixture regions following (Pool et al. 2012; Lack et al. 2015). At 100kb-scale, ρ_{LD} estimated across the sequences of the RG and ZI populations show a Spearman's r_s of 0.76 ($P < 1 \times 10^{-16}$).

To investigate whether the main trends observed in this study were driven by estimates of the population parameter ρ_{LD} that, at least partially, reflect differences in N_e rather than in recombination rate r, we normalized ρ_{LD} by estimates of silent nucleotide diversity (π_{sil}; $\pi_{sil} = 4N_e\mu$ in autosomes and $3N_e\mu$ in the *X* chromosome, where μ is the mutation rate per bp and generation). To obtain π_{sil}, we estimated pairwise nucleotide variation per site at intergenic sequences, introns and 4-fold degenerate sites in coding regions from sequences of the RG population. Finally, we studied the influence of variable motif presence across the genome on the ratio $\rho_{LD/}\pi_{sil}$, which should be independent of N_e when both ρ_{LD} and π_{sil} are estimated from the same population and at the same genomic scale.

Open Chromatin Profiles and Transcriptional Activity in Early Female Meiosis

We used high-resolution open chromatin profiles in *D. melanogaster* from four differences samples and with two different methods. High-magnitude DNase I hypersensitive sites (DHS) from three different cell lines were obtained from http://compbio.hms.harvard.edu/kharchenko-et-al-nature-2011 (Kharchenko et al. 2011; Thomas et al. 2011). The three cell lines investigated are: S2 (S2-DRSC; derived from late male embryonic tissues, stages 16–17), BG3 (ML-DmBG3-c2; derived from the central nervous system of male third instar larvae), and Kc167 (derived from disaggregated young embryos 8–12 h old). We also obtained information on accessible chromatin regions in the eye-antennal imaginal disc generated using the recently developed ATAC methodology [(Davie et al. 2015); NCBI GEO samples GSM1426254- GSM1426256]. ATAC [Assay for Transposase Accessible Chromatin; (Buenrostro et al. 2015; Sos et al. 2016)] is more sensitive and robust than DNase-seq for identifying sites of high chromatin accessibility and thus can provide information more relevant for *in vivo* DSB (Davie et al. 2015). Transcriptome information during early female meiosis in *D. melanogaster* was generated by Adrian and Comeron (2013) and RNA-seq data are available at NCBI SRA SRP032523.

Analyses, Model Generation and Attribute Selection

We applied LASSO (Tibshirani 1996; Hastie et al. 2009) as implemented in the WEKA v3.6.1 software package [(Hall et al. 2009); http://www.cs.waikato.ac.nz/ml/weka/]. LASSO is a data mining technique that favors solutions with fewer parameter values under a *linear* model, simultaneously performing variable selection and simplifying model interpretation. The intensity of regularization (or shrinkage) within LASSO is controlled by the regularization/shrinkage parameter (λ). Unless noted, we used a λ that minimizes the cross validated mean squared error plus 0.5 standard error. Cross-validation (CV), which we used within LASSO, RF, and MARS analyses (see below), simulates the process of separately developing a model on one set of data and predicting for a test set of data not used in developing the model, with all aspects of the model development process repeated for each loop of the cross-validation.

We also utilized the WEKA implementation of Random Forests (RF) for classification. RF is a non-parametric approach useful for detecting associations when there are large numbers of predictor variables with the possibility that each variable has relatively weak effects (Breiman 2001; Banfield et al. 2007). Briefly, RF classification constructs a collection of many independent decision trees, sampling both the data and attributes randomly with replacement. The remaining, unused data is classified using the collection of trees, with the classification of each item being based upon the result mode of the RF. Here, we generated 1,000 trees of unrestricted depth with Log_2(Attribute Number) + 1 random attributes in each individual tree. RF models were evaluated using 10-fold cross validation (CV), which involves splitting the complete dataset into ten equal sets and training on nine sets while testing on the remaining set—this process is then repeated ten times to obtain an average accuracy for each class. An exception to the 10-CV evaluation method was applied when studying how a model trained with data from an individual chromosomal arm performed when applied to other chromosomal arms, in which cases we used the full data from the chromosome used as training set and the full data from the chromosome used to test the model.

Each model was tested versus a ZeroR null model which classifies all instances solely based on the majority (mode) class. In all cases, the RF model performed significantly better at classification (two-tailed *t*-test, $P < 0.05$) than the null model unless otherwise noted. In order to select the best features for use in model generation, we ranked all features by the information gain criterion implemented in WEKA. Information gain is the measure of the contribution of a particular feature to the model. When calculating probabilities for RF models, we performed 250,000 whole-genome randomizations per model. For each randomization (replicate), RF modeling was limited to 150 trees using 10-fold cross validation in order to speedup computation at the expense of increased variance, and therefore generates conservative estimates. Conversely, when testing RF on the fourth chromosome, we used the complete genomic data from 2L, 2R, 3L, 3R, and the *X* for training and tested on the whole chromosome as well as on the thirteen adjacent 100-kb regions of this small chromosome. Tests on the fourth chromosome were repeated 20,000 times with 150 trees on random seeds in order to evaluate reliability.

We applied multivariate adaptive regression splines (MARS) (Friedman 1991; Friedman and Roosen 1995; Hastie et al. 2009) using the software suite Salford Predictive Modeler (v.7) from Salford Systems (http://www.salford-systems.com). MARS is a form of regression analysis that splits predictive variables into several intervals, allows potential non-linear relationships over different intervals (basis functions) and combines individual models as a final quantitative and predictive model. Importantly, MARS allows for any degree of interaction between variables. The quality of MARS models can be ascertained using the generalized cross-validation (GCV) criterion, with small GCV scores being indicative of superior model fit (Craven and Wahba 1979). Because the number of data points is not very large ($n = 1{,}191$ non-overlapping 100-kb regions), we avoided portioning the data into training and test samples and, instead, we initially applied cross-validation and the MARS legacy modes to estimate the optimal model and its performance (Friedman 1991). Note, however, that under the legacy mode MARS builds a sequence of models using all available data and can overestimate the performance of the model. We, thus, show MARS results based on 10-fold cross-validation to train and test classifiers unless specifically noted. We use the notation R^2_{CV} to describe the MARS estimate of how well this model would perform on new data when using the 10-fold cross validation mode. In all cases, a

10-fold cross-validation mode generates smaller (more conservative) R^2_{CV} than when a legacy mode is applied ($R^2_{Legacy} > R^2_{CV}$). Alternative estimates of model quality such standard R^2 using the whole dataset can be subject to substantial overfitting and are not reported (we obtained $R^2 > R^2_{CV}$ in all cases). Finally, variable importance under MARS was measured by the Gini index (Breiman et al. 1984).

Supplementary Material

Supplementary tables S1–S4 and figures S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Ana Llopart for comments and suggestions on the manuscript. This study was supported by The National Science Foundation (NSF, Division of Environmental Biology) Grant DEB1354921.

Literature Cited

Adrian AB, Comeron JM. 2013. The *Drosophila* early ovarian transcriptome provides insight to the molecular causes of recombination rate variation across genomes. BMC Genomics 14:794.

Aymard F, et al. 2014. Transcriptionally active chromatin recruits homologous recombination at DNA double-strand breaks. Nat Struct Mol Biol. 21:366–374.

Bailey TL, et al. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37:W202–W208.

Banfield RE, Hall LO, Bowyer KW, Kegelmeyer WP. 2007. A comparison of decision tree ensemble creation techniques. IEEE Trans Pattern Anal Mach Intell 29:173–180.

Baudat F, et al. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327:836–840.

Bauer E, et al. 2013. Intraspecific variation of recombination rate in maize. Genome Biol. 14:R103.

Beadle GW. 1932. A possible influence of the spindle fibre on crossing-over in *Drosophila.* Proc Natl Acad Sci U S A. 18:160–165.

Berchowitz LE, Hanlon SE, Lieb JD, Copenhaver GP. 2009. A positive but complex association between meiotic double-strand break hotspots and open chromatin in *Saccharomyces cerevisiae.* Genome Res. 19:2245–2257.

Bessoltane N, Toffano-Nioche C, Solignac M, Mougel F. 2012. Fine scale analysis of crossover and non-crossover and detection of recombination sequence motifs in the honeybee (*Apis mellifera*). PLoS One 7:e36229.

Billings T, et al. 2013. DNA binding specificities of the long zinc-finger recombination protein PRDM9. Genome Biol. 14:R35.

Borde V, de Massy B. 2013. Programmed induction of DNA double strand breaks during meiosis: setting up communication between DNA and the chromosome structure. Curr Opin Genet Dev. 23:147–155.

Breiman L. 2001. Random forests. Mach Learn 45:5–32.

Breiman L, Friedman J, Olshen R, Stone C. 1984. Classification and regression trees. Boca Raton: CRC Press.

Brooks LD. 1988. The evolution of recombination rates. In: Michod , R.E., Levin, B.R., editors. The evolution of sex. Sunderland (MA): Sinauer Associates. p. 87–105.

Brooks LD, Marks RW. 1986. The organization of genetic variation for recombination in *Drosophila melanogaster.* Genetics 114:525–547.

Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. 2015. ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol. 109:21 29 21–29.

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

Charlesworth B, Campos JL. 2014. The relations between recombination rate and patterns of molecular variation and evolution in *Drosophila.* Annu Rev Genet. 48:383–403.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303.

Comeron JM. 2014. Background selection as baseline for nucleotide variation across the *Drosophila* genome. PLoS Genet. 10:e1004434.

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

Coop G, Myers SR. 2007. Live hot, die young: transmission distortion in recombination hotspots. PLoS Genet. 3:e35.

Coop G, Przeworski M. 2007. An evolutionary view of human recombination. Nat Rev Genet. 8:23–34.

Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. 2008. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. Science 319:1395–1398.

Craven P, Wahba G. 1979. Smoothing noisy data with spline functions. Numer Math 31:377–403.

Cromie GA, et al. 2007. A discrete class of intergenic DNA dictates meiotic DNA break hotspots in fission yeast. PLoS Genet. 3:e141.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res. 14:1188–1190.

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

Davie K, et al. 2015. Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. PLoS Genet. 11:e1004994.

de Castro E, et al. 2012. Nucleosomal organization of replication origins and meiotic recombination hotspots in fission yeast. EMBO J. 31:124–137.

Dlakic M, Harrington RE. 1998. Unconventional helical phasing of repetitive DNA motifs reveals their relative bending contributions. Nucleic Acids Res. 26:4274–4279.

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

Dumont BL, et al. 2011. Extensive recombination rate variation in the house mouse species complex inferred from genetic linkage maps. Genome Res. 21:114–125.

Fan Q-Q, Petes TD. 1996. Relationship between nuclease-hypersensitive sites and meiotic recombination hot spot activity at the HIS4 locus of *Saccharomyces cerevisiae.* Mol Cell Biol. 16:2037–2043.

Fiston-Lavier AS, Singh ND, Lipatov M, Petrov DA. 2010. *Drosophila melanogaster* recombination rate calculator. Gene 463:18–20.

Fledel-Alon A, et al. 2011. Variation in human recombination rates and its genetic determinants. PLoS One 6:e20321.

Fowler KR, Sasaki M, Milman N, Keeney S, Smith GR. 2014. Evolutionarily diverse determinants of meiotic DNA break and recombination landscapes across the genome. Genome Res. 24:1650–1664.

Friedman JH. 1991. Multivariate adaptive regression splines. Ann Stat 1–67.

Friedman JH, Roosen CB. 1995. An introduction to multivariate adaptive regression splines. Stat Methods Med Res. 4:197–217.

Garner MM, Felsenfeld G. 1987. Effect of Z-DNA on nucleosome placement. J Mol Biol. 196:581–590.

Gerton JL, et al. 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. Proc Natl Acad Sci U S A. 97:11383–11390.

Gillespie JH. 2000. Genetic drift in an infinite population. The pseudohitchhiking model. Genetics 155:909–919.

Gossmann TI, Santure AW, Sheldon BC, Slate J, Zeng K. 2014. Highly variable recombinational landscape modulates efficacy of natural selection in birds. Genome Biol Evol. 6:2061–2075.

Graffelman J, Balding DJ, Gonzalez-Neira A, Bertranpetit J. 2007. Variation in estimated recombination rates across human populations. Hum Genet. 122:301–310.

Hall M, et al. 2009. The WEKA data mining software: an update. SIGKDD Explor. Newsl 11:10–18.

Hastie T, et al. 2009. The elements of statistical learning. New York: Springer.

Heil CS, Noor MA. 2012. Zinc finger binding motifs do not explain recombination rate variation within or between species of *Drosophila.* PLoS One 7:e45055.

Herbert A, Rich A. 1996. The biology of left-handed Z-DNA. J Biol Chem. 271:11595–11598.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genet Res. 8:269–294.

Hinch AG, et al. 2011. The landscape of recombination in African Americans. Nature 476:170–175.

Hirota K, Steiner WW, Shibata T, Ohta K. 2007. Multiple modes of chromatin configuration at natural meiotic recombination hot spots in fission yeast. Eukaryot Cell 6:2072–2080.

Hizver J, Rozenberg H, Frolow F, Rabinovich D, Shakked Z. 2001. DNA bending by an adenine–thymine tract and its role in gene regulation. Proc Natl Acad Sci U S A. 98:8490–8495.

Hochman B. 1973. The fourth chromosome of *Drosophila melanogaster*. In: Novitski E, Ashburner M, editors. The genetics and biology of Drosophila. London, New York, San Francisco: Academic Press. p. 903–928.

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

Hussin J, Roy-Gagnon MH, Gendron R, Andelfinger G, Awadalla P. 2011. Age-dependent recombination rates in human pedigrees. PLoS Genet. 7:e1002251.

Jiang H, et al. 2011. High recombination rates and hotspots in a *Plasmodium falciparum* genetic cross. Genome Biol. 12:R33.

Keeney S. 2001. Mechanism and control of meiotic recombination initiation. Curr Topics Dev Biol. 52:1–53.

Kharchenko PV, et al. 2011. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster.* Nature 471:480–485.

Kim S, et al. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana.* Nat Genet. 39:1151–1155.

Kleckner N. 2006. Chiasma formation: chromatin/axis interplay and the role(s) of the synaptonemal complex. Chromosoma 115:175–194.

Koehler KE, et al. 1996. Spontaneous X chromosome MI and MII nondisjunction events in *Drosophila melanogaster* oocytes have different recombinational histories. Nat Genet. 14:406–414.

Kong A, et al. 2002. A high-resolution recombination map of the human genome. Nat Genet. 31:241–247.

Kong A, et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. Nature 467:1099–1103.

Koszul R, Kim KP, Prentiss M, Kleckner N, Kameoka S. 2008. Meiotic chromosomes move by linkage to dynamic actin cables with transduction of force through the nuclear envelope. Cell 133:1188–1201.

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

Lack JB, et al. 2015. The *Drosophila* genome nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. Genetics 199:1229–1241.

Lee JW, Lee JB, Park M, Song SH. 2005. An extensive comparison of recent classification tools applied to microarray data. Comput Stat Data Anal 48:869–885.

Lindsley DL, Zimm GG. 1992. The genome of Drosophila melanogaster. San Diego (CA): Academic Press.

Liu H, et al. 2015. Causes and consequences of crossing-over evidenced via a high-resolution recombinational landscape of the honey bee. Genome Biol. 16:15.

Llopart A. 2015. Parallel faster-x evolution of gene expression and protein sequences in *Drosophila*: beyond differences in expression properties and protein interactions. PLoS One 10:e0116829.

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454:479–485.

McGaugh SE, et al. 2012. Recombination modulates how selection affects linked sites in *Drosophila.* PLoS Biol. 10:e1001422.

Meisel RP, Connallon T. 2013. The faster-X effect: integrating theory and data. Trends Genet. 29:537–544.

Miller DE, et al. 2012. A whole-chromosome analysis of meiotic recombination in *Drosophila melanogaster.* G3 (Bethesda) 2:249–260.

Mirouze M, et al. 2012. Loss of DNA methylation affects the recombination landscape in *Arabidopsis.* Proc Natl Acad Sci U S A. 109: 5880–5885.

Morton NE, Rao DC, Yee S. 1976. An inferred chiasma map of *Drosophila melanogaster.* Heredity (Edinb) 37:405–411.

Muñoz-Fuentes V, Di Rienzo A, Vilà C. 2011. Prdm9, a major determinant of meiotic recombination hotspots, is not functional in dogs and their wild relatives, wolves and coyotes. PLoS One 6:e25498.

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

Myers S, et al. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. Nat Genet. 40:1124–1129.

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

Neumann R, Jeffreys AJ. 2006. Polymorphism in the activity of human crossover hotspots independent of local DNA sequence variation. Hum Mol Genet. 15:1401–1411.

Ohta K, Shibata T, Nicolas A. 1994. Changes in chromatin structure at recombination initiation sites during yeast meiosis. EMBO J. 13:5754–5763.

Oliver PL, et al. 2009. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS Genet. 5:e1000753.

Pan J, Keeney S. 2007. Molecular cartography: mapping the landscape of meiotic recombination. PLoS Biol. 5:e333.

Pan J, et al. 2011. A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. Cell 144:719–731.

Parsons PA. 1988. Evolutionary rates: effects of stress upon recombination. Biol J Linn Soc 35:49–68.

Parvanov ED, Petkov PM, Paigen K. 2010. Prdm9 controls activation of mammalian recombination hotspots. Science 327:835.

Parvinen M, Soderstrom KO. 1976. Chromosome rotation and formation of synapsis. Nature 260:534–535.

Perez-Martin J, de Lorenzo V. 1997. Clues and consequences of DNA bending in transcription. Annu Rev Microbiol. 51:593–628.

Petes TD. 2001. Meiotic recombination hot spots and cold spots. Nat Rev Genet. 2:360–369.

Pool J, et al. 2012. Population Genomics of sub-saharan *Drosophila melanogaster*: African diversity and non-African admixture. PLoS Genet. 8:e1003080.

Pool JE. 2015. The mosaic ancestry of the Drosophila Genetic Reference Panel and the *D. melanogaster* reference genome reveals a network of epistatic fitness interactions. Mol Biol Evol. 32:3236–3251.

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

Ross JA, et al. 2011. *Caenorhabditis briggsae* recombinant inbred line genotypes reveal inter-strain incompatibility and the evolution of recombination. PLoS Genet. 7:e1002174.

Segal E, et al. 2006. A genomic code for nucleosome positioning. Nature 442:772–778.

Ségurel L, Leffler EM, Przeworski M. 2011. The case of the fickle fingers: how the PRDM9 zinc finger protein specifies meiotic recombination hotspots in humans. PLoS Biol. 9:e1001211.

Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the *Drosophila* genome? PLoS Genet. 5:e1000495.

Shibuya H, Ishiguro K, Watanabe Y. 2014. The TRF1-binding protein TERB1 promotes chromosome movement and telomere rigidity in meiosis. Nat Cell Biol. 16:145–156.

Shimizu M, Mori T, Sakurai T, Shindo H. 2000. Destabilization of nucleosomes by an unusual DNA conformation adopted by poly(dA) small middle dotpoly(dT) tracts in vivo. EMBO J. 19:3358–3365.

Singh ND, et al. 2015. Fruit flies diversify their offspring in response to parasite infection. Science 349:747–750.

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 194:375–387.

Slatkin M, Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555–562.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res. 23:23–35.

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

Smukowski Heil CS, Ellison C, Dubin M, Noor MA. 2015. Recombining without Hotspots: a comprehensive evolutionary portrait of recombination in two closely related species of *Drosophila.* Genome Biol Evol. 7:2829–2842.

Soriano I, Quintales L, Antequera F. 2013. Clustered regulatory elements at nucleosome-depleted regions punctuate a constant nucleosomal landscape in *Schizosaccharomyces pombe.* BMC Genomics 14:813.

Sos BC, et al. 2016. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. Genome Biol. 17:20.

Steiner WW, Steiner EM, Girvin AR, Plewik LE. 2009. Novel nucleotide sequence motifs that produce hotspots of meiotic recombination in *Schizosaccharomyces pombe.* Genetics 182:459–469.

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. 12:530–532.

Stevison LS, et al. 2015. The time-scale of recombination rate evolution in great apes. Mol Biol Evol. 33:928–945.

Struhl K, Segal E. 2013. Determinants of nucleosome positioning. Nat Struct Mol Biol. 20:267–273.

Sun FL, et al. 2000. The fourth chromosome of *Drosophila melanogaster*: interspersed euchromatic and heterochromatic domains. Proc Natl Acad Sci U S A. 97:5340–5345.

Thomas S, et al. 2011. Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development. Genome Biol. 12:R43.

Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. J R Stat Soc B Methodol 58:267–288.

Travers AA. 1990. Why bend DNA? Cell 60:177–180.

Treco D, Arnheim N. 1986. The evolutionarily conserved repetitive sequence d(TG.AC)n promotes reciprocal exchange and generates unusual recombinant tetrads during yeast meiosis. Mol Cell Biol. 6:3934–3947.

Wright S. 1938. Size of population and breeding in relation to evolution. Science 87:430–431.

Wu TC, Lichten M. 1994. Meiosis-induced double-strand break sites determined by yeast chromatin structure. Science 263:515–518.

Yamada T, Ohta K. 2013. Initiation of meiotic recombination in chromatin structure. J Biochem. 154:107–114.

Yelina NE, et al. 2012. Epigenetic remodeling of meiotic crossover frequency in *Arabidopsis thaliana* DNA methyltransferase mutants. PLoS Genet. 8:e1002844.

Associate editor: Dr Maria Costantini

Highlights editor: George Zhang