

Statistical inference of recombination-inducing genic features in *Drosophila melanogaster*

JONAS MUELLER

May 25, 2013

Undergraduate Honors Thesis
Department of Statistics
University of California, Berkeley
Supervisor: Yun S. Song
Professor of Statistics and EECS

ABSTRACT

Building on the findings of Chan et al. [1], this work applies a variety of computational methods to further our limited understanding of recombination-initiating processes in *Drosophila melanogaster*. Using recently produced high resolution recombination maps for two different *D. melanogaster* populations, we identify recombination hotspots (regions of extremely elevated recombination activity) across the genome for each population. In addition to characterizing genic features which differ in the hotspot regions compared with the rest of the genome, we present a number of DNA sequence motifs that are significantly overrepresented in the hotspots, suggesting a possible recombination-inducing role played by these patterns. Subsequently, we conduct a wavelet coherence analysis between genic features and the recombination maps, discovering significant relationships between recombination rates and nucleosome, gene, and transposable element locations. We also find that the locations of the GCCAATTT motif, a binding site of the homeobrain transcription factor, as well as the 4-mer GCCA, which comprises the core of numerous Polycomb-group transcription factor binding sites, are strongly associated with elevated local recombination rates in numerous regions across the genomes for both *D. melanogaster* populations.

INTRODUCTION

Genetic recombination is the process by which a double-stranded DNA molecule breaks and joins with another. Chromosomal crossover, for example, is a recombination event in which a pair of homologous parental chromosomes exchange regions of their DNA to form new chromosomes that will produce genotypically and phenotypically different offspring. Serving as a significant source of genetic variation in sexual organisms, this process is of fundamental importance as it is responsible for not only the diversity in life on our planet, but also for numerous diseases which occur when recombination events disturb critical regions of the genome.

Although recombination events can occur at almost any location in the genome, it has been shown that these events are far from uniformly distributed and the genomes of many species contain narrow regions, ~ 2 kilobases (kb) long, with significantly elevated rates of recombination [2, 3]. These loci are referred to as recombination hotspots, and their existence suggests that recombination is, to some extent, associated with local features in the underlying DNA sequence. Consequently the identification of genic features near recombination hotspots has recently drawn significant attention and several important discoveries have been made.

Recombination between homologous chromosomes in all higher eukaryotes is thought to be initiated by the formation of a double stranded break (DSB) in the DNA. It has been observed that such a break is the product of enzymatic activity induced by certain protein complexes, and the chromosomes recombine if the DSB is followed by the repair and ligation of the two

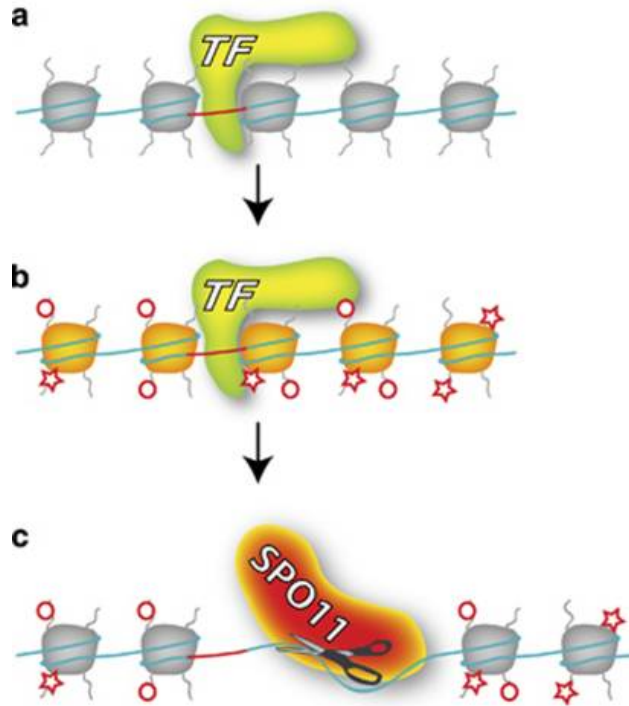


Figure 1: A simplified representation of Wahls and Davidson's pan-eukaryotic model for recombination positioning. **(a)** Consensus DNA recognition sites (dark segment) within recombination hotspots are recognized by a transcription factor (TF), such as *PRDM9* in mammals or *Atf1-Pcr1* in yeast. Identifiable sequence motifs near hotspots fall into many classes, each presumably recognized by a separate TF. **(b)** In yeast, TFs recruit histone modification enzymes that add, or subtract, methyl, acetyl or ubiquitin marks (circles or stars), for example, to adjacent nucleosomes. In mammals, the hotspot-recognizing *PRDM9* is itself a H3K4 trimethylase. **(c)** These histone modifications then recruit further elements of the recombination machinery, including *Spo11*, presumably to more open chromatin. This catalyses DSBs in the vicinity of the hotspot. (Figure and caption from Goodstadt and Ponting [14])

different DNA-strands by specialized meiotic DNA repair proteins [4]. It is generally believed that the presence of an open chromatin structure is one important feature of recombination hotspots [5,6]. This structure can be acquired in many ways, including binding of transcription factors [7], intrinsic sequence features and epigenetic markers such as histone modifications [8] (see Figure 1 for an example). One recent study showed that the transcription factor *Atf1* activates meiotic recombination hotspots in *Schizosaccharomyces pombe* [9]. The protein contains a bZIP domain and together with its heterodimer partner *Pcr1*, it binds specifically to a 7 base pair (bp) DNA sequence motif. In the human genome, a 13 bp motif CCNCCNTNNCCNC has been shown to be associated with over 40% of chromosomal crossover events [10]. This motif is a binding site for the zinc-finger protein *PRDM9*, which has been shown to initiate DSB formation in the vicinity of its binding location across the genomes of humans and other mammals, such as primates and mice [11–13].

Although the drivers of recombination have been extensively investigated in yeast, humans, and other mammals, these mechanisms are less understood in flies. Goodstadt and Ponting have pointed out that the question of whether control of recombination mechanisms is conserved among diverse eukaryotes remains an important open problem [14]. The main focus of this work is to characterize sequence motifs and other genic features which are correlated with elevated rates

of recombination in *Drosophila melanogaster*, a model fly organism commonly studied in genetic experiments. There are numerous features of recombination in *Drosophila* which distinguish the process from recombination in other organisms like yeast, humans, mice, and plants, and it has been proposed that *Drosophila* possess a different recombination initiation mechanism than these other organisms [15]. For example, *Drosophila* lack genes known to be integral to the recombination process in other organisms [16] and Heil and Noor recently failed to find evidence of any *Drosophila* homolog of *PRDM9* involved in initiating recombination events [15].

Previous experimental studies have identified none to few *Drosophila* hotspots and found that the spike in recombination rates at these regions is generally far milder than the ten-to-hundred-fold rate increases observed at mammal, plant, and yeast hotspots [17–20]. In a pedigree study, Miller et al. localized fifteen crossover events on the *D. melanogaster* X chromosome, finding the 7 bp motif GTGGAAA significantly enriched in the vicinity of these crossovers [21]. Comeron et al. recently conducted large scale *D. melanogaster* interbreeding experiments to obtain estimates of genome-wide crossover rates and they also found a number of motifs associated with elevated recombination rates. However, all past examinations of genome-wide recombination in *D. melanogaster* have been limited by low resolution estimates of rates (on the order of hundreds of kilobases), from which it is difficult to pinpoint specific local genic features that influence jumps in recombination rate. To infer maps of recombination rates across the genome, computational methods based on the linkage disequilibrium (LD: a measure of the dependence between genotypes at different loci) in a sample have become widely adopted [1, 17, 22]. In this work, we first identify recombination hotspots across the *D. melanogaster* autosomal genome from recently produced fine-scale LD-based recombination maps, then employ a variety of methods to search for sequence motifs overrepresented in these hotspots, and finally investigate which candidate motifs and other genic features exhibit local correlations with estimated recombination rates in various regions across the genome.

DATA

For our analysis, we employ the pair of genome-wide recombination maps recently constructed by Chan, Jenkins, and Song for two *D. melanogaster* populations, one from Raleigh, USA (RAL) and the other from Gikongoro, Rwanda (RG) [1]. To obtain very fine-scale LD-based estimates of recombination rates throughout the *Drosophila* genomes, the authors used a reversible-jump Markov Chain Monte Carlo method called LDhelmet which utilizes recent theoretical advances in asymptotic sampling distributions to improve the computation of likelihoods in the population genetic model adopted for these populations. The recombination rates were estimated by applying LDhelmet to the RAL dataset, consisting of 37 genomes from inbred lines sequenced at a coverage of $\geq 10\times$ by the Drosophila Population Genomics Project (Release 1.0 from DPGP, www.dpgp.org/), and the RG dataset: 22 genomes from haploid embryos sequenced at a coverage of $\geq 25\times$ by the Drosophila Population Genomics Project 2 (Release 2.0 from DPGP2, www.dpgp.org/dpgp2/DPGP2.html) [23].

After first dividing the data into overlapping blocks of 4,400 SNPs each (with 200 SNPs overlapping between each pair of adjacent blocks), LDhelmet was run for 3,000,000 iterations on each block (after 300,000 iterations of burn-in). The final recombination map for each chromosome arm was then produced by stitching the blocks together after removing the 200 overlapping SNPs between each pair of blocks [1]. The extremely high SNP density in the data (~ 1 SNP per 38 bp in the 22 samples of the RG dataset) allows recombination rate variation to be localized to very fine scales, and in their work, Chan et al. find that the resulting recombination map is highly correlated with the experimental genetic map produced by Singh et al. [18] and those hosted at the Flybase website (www.flybase.org [24]). They also demonstrate that LDhelmet exhibits superior perfor-

mance over the widely used LDhat recombination-rate estimation method [22] in the presence of natural selection.

LDhelmet requires a prior distribution on the number of change points in the recombination map, which is determined by a user-defined parameter called the *block penalty* (BP). Responsible for controlling the degree of variation in the estimated recombination rates, the block penalty was fixed at a conservative value of 50 by Chan et al. (where higher block penalties result in estimation of smoother recombination maps), and their resulting recombination map estimates are publicly available at: <http://sourceforge.net/projects/ldhelmet/> [1]. While a high BP of 50 strongly limits false positive inference of hotspots, it also results in a severe decrease in our power to identify regions of elevated recombination due to the restricted variation in the estimated recombination map. Because Myers et al. successfully identified overrepresentation of the *PRDM9*-binding motif in a large number of hotspots estimated from a human recombination map constructed using a much lower BP of 5 (albeit using the LDhat method [22] rather than LDhelmet) [17], we also reapply LDhelmet to the *Drosophila* sequences with the less conservative block penalty 10 (keeping all other parameters the same as those specified in [1]). A comparison between the recombination rate estimates under these two choices of block penalty shows that besides exhibiting a greater degree of variation, the BP-10 LDhelmet results do not differ much from the BP-50 estimates (see Supplementary Information).

Because sex chromosomes are subject to markedly different mechanisms of recombination from the rest of the genome in *D. melanogaster* and numerous other dioecious organisms [25], we restrict our analysis to the two major autosomal chromosomes of the sequenced *Drosophila* genomes (also omitting the minuscule fourth chromosome due to its lack of recombination activity [26]). This leaves us with four major regions of the *D. melanogaster* genome in which we study the estimated recombination rates of the RAL/RG populations: arm 2L (the 23 Mega-base-pair (Mb) left arm of chromosome 2), arm 2R (the 21 Mb right arm of chromosome 2), arm 3L (the 24.5 Mb left arm of chromosome 3), and arm 3R (the 28 Mb right arm of chromosome 3). To investigate the relationships between recombination rates and various genic features, we also make use of the Flybase *D. melanogaster* genome annotations (release 5.45, <http://www.flybase.org> [24]) to identify the location of features with respect to the recombination maps.

RESULTS

From the RAL/RG recombination maps estimated using block penalties 10 and 50, we first identify hotspots of recombination in each population. As Chan et al. found only moderate correlation between the RAL and RG estimated recombination rates, we do not attempt to merge the two maps, choosing to identify hotspots separately for each population as they did in their work [1]. Following the methodology of Myers et al., we then randomly select a corresponding “coldspot” control region for each hotspot and search for DNA motifs significantly overrepresented in the vicinity of the hotspot regions compared with the coldspots [10]. For each hotspot, a putative coldspot sequence of identical length is randomly selected from a 100 kb region on the same chromosome arm around the hotspot, and the sequence is only accepted as the corresponding coldspot once it meets criteria I-IV:

- I. Average recombination rate in the putative coldspot region does not exceed the chromosome-arm-wide average rate
- II. Putative coldspot region does not overlap with any hotspot
- III. Proportion of degenerate (IUPAC code: “N”) nucleotides in putative coldspot does not exceed the proportion of degenerate nucleotides in the corresponding hotspot

IV. GC content, exon content, and diversity of hotspot region are matched in putative coldspot

where GC content (the fraction of sequenced nucleotides called as G or C), exon content (the fraction of region with exotic annotation in Flybase), and diversity (the average fraction, across pairs of samples within the population, of sites that differ between each pair, out of the number of sites for which both samples have data) are matched between hot/cold-spots because these features were found by Chan et al. to be correlated with recombination rates and thus are potentially confounding factors in our search for motifs that explain the difference in recombination rates between the hot/cold-spots [1].

RECOMBINATION HOTSPOTS

Following Chan et al., we identify hotspots as regions of size ≥ 500 bp in which the recombination rate exceeds ten times the chromosome-arm-wide average [1]. 500 bp was chosen as the minimum required length of a hotspot because narrow peaks in estimated recombination rates can occasionally occur as spurious artifacts of LDhelmet’s reversible-jump Markov Chain Monte Carlo procedure. However, we do not adopt the highly conservative filter of Chan et al., which involved separately applying the sequenceLDhot method [27] and only retaining those hotspots which overlap with a region identified by sequenceLDhot. Our decision to skip the sequenceLDhot filtering is based on a desire to retain as much power as possible in our motif search and the method’s numerous disadvantages, including its lack of accuracy [27] as well as its biased preference for spots in which the local background recombination rate is already higher than the chromosome-arm-wide mean rate [1].

The significantly larger number of hotspots in the two right-most columns of Table 1 reflects the greater variation allowed in the recombination maps produced under block penalty 10. We note that every hotspot region in the BP-50 recombination maps is represented in the BP-10 maps, and because the increase in hotspot number provides a wealth of information in our search for hotspot-inducing sequence motifs, we focus the remainder of our analysis on the BP-10 recombination maps (any mention of recombination maps/hotspots with unspecified block penalty is hereafter assumed to refer to the estimates produced under block penalty 10).

Although Chan et al. demonstrated a strong global correlation between the two BP-50 recombination maps in a wavelet coherence analysis, they found little correlation between the estimated recombination rates at finer scales, which they suggest is partly explained by biological differences between the populations [1]. Due to the lack of correlation at finer scales, we find in Table 2, that very few of the inferred hotspot regions are present in both maps. However this discrepancy is not as severe as it appears. For the vast majority of hotspots in the RAL group, the corresponding region of the RG recombination map does in fact contain a narrow peak in estimated recombination rates well above 10 times the chromosome-wide average rate, but most of these spikes are not

Arm	RAL (BP=50)	RG (BP=50)	RAL (BP=10)	RG (BP=10)
2L	2	1	32	38
2R	3	2	39	31
3L	3	1	58	30
3R	10	8	58	55
Total:	18	12	187	154

Table 1: The number of hotspots identified from the recombination map of the individual chromosome arms in each population, produced using one of two choices of block penalty (indicated in parentheses). The bottom row contains the total number of hotspots across all four chromosome arms.

Arm	Hotspots (RAL)	Hotspots (RG)	Shared	Total	Hotspots / Mb
2L	32	38	0	70	3.04
2R	39	31	1	70	3.31
3L	58	30	5	88	3.59
3R	58	55	6	113	4.05

Table 2: The number of hotspots identified in each chromosome arm for the RAL and RG populations (under block penalty 10). The Shared column counts the hotspots in the RAL group for which there a corresponding hotspot in the RG group within 1 kb of the same location, the Total column gives the total number of hotspots across both groups, and in the last column, these totals are scaled by the the length of the chromosome arm (in Megabases).

identified as hotspots because they fail to meet the 500 bp length requirement. We compute the estimated mean RG recombination rate across all regions identified as hotspots from the RAL map and find that it is 22 times the chromosome-wide mean RG rate in chromosome arm 2L, 24 times the chromosome-wide average in chromosome arm 2R, 19 in 3L, and 51 in 3R.

Figure 2 illustrates that under our hotspot definition, the majority of regions we identify from the BP-10 recombination maps are just under a couple of kb in length, agreeing with the general consensus of hotspot size across a number of organisms [2,3]. There is one enormously large hotspot region identified on chromosome arm 3R in the RG population, whose 30 kb size is likely overestimated, but since we lack information to narrow the search for potential recombination-driving motifs that might lie in this region of inferred recombination abundance, we leave this hotspot unmanipulated and extract a coldspot of equally large size. It is unlikely that this single lengthy hot/cold-spot pair will significantly affect our search for sequence motifs involved in determining the location of recombination hotspots, and we do not wish to omit this region from our analysis due to its significantly elevated estimated recombination rate. Figure 3 depicts the locations of all hot/cold spot regions we identify along the recombination map for each population.

To obtain hotspot sequences corresponding to the hotspot regions, we consider two approaches. In the first approach, we choose one of the sequenced individuals in each population to serve as a reference and extract sequences from this sample’s chromosome arms. However, due to the high SNP density in *D. melanogaster*, this method leads to a loss of quite a bit of sequence information contained within the numerous nucleotide variations between our samples. To retain this information, we can instead extract sequences corresponding to hotspot locations from all individuals in each population. After obtaining coldspot sequences from all individuals via the previously described criteria, we can then identify motifs overrepresented in the hotspot sequences. Unfortunately, this approach tends to favor motifs in subregions of the hotspots with low SNP density between the samples, since in such an area, the sequences of all individuals share many of the same motifs, while the presence of the motif in the corresponding coldspots may be masked by nucleotide mismatches if it happens to be located in a SNP-rich region. In this case, it will seem that this motif is significantly less present in the coldspots, even if it has no role in determining recombination. Because we deem the potential SNP-density bias which can result from this second approach more problematic than the loss of information accompanying the first method, we only extract hot/cold-spot sequences from a single sample in each population. For this task, RAL-707 and RG38N are the samples chosen from the RAL and RG groups, respectively, because these are the individuals in each dataset for which the hotspot regions have been the most thoroughly sequenced. Due to the inherent uncertainty regarding the start/end-points of the hotspot regions as well where potential motifs of interest are located within the hotspots, we extend each hotspot region by 500 bp on both ends before identifying the sequence associated with the hotspot.

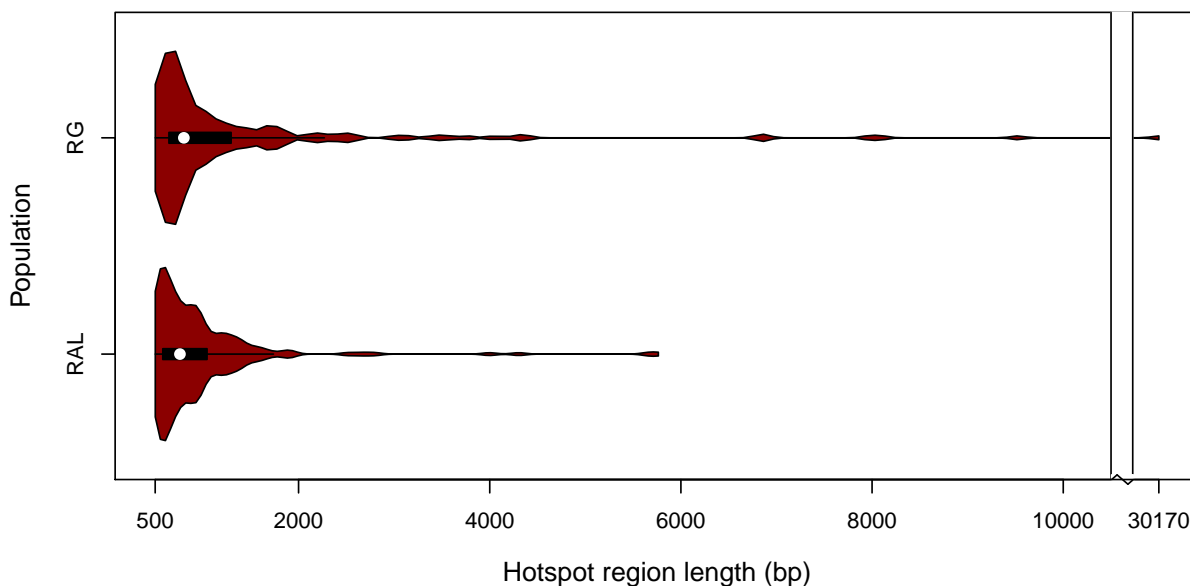


Figure 2: Violin plots depicting the length distribution of the 187 and 154 hotspot regions identified from the recombination maps of the RAL and RG populations, respectively. The width of each plot reflects a kernel density estimate of the distribution of the hotspot region sizes, and the black rectangles are box plots of the hotspot lengths with the medians marked in white. To facilitate visualization of the distributions, the outlying 30,170 bp hotspot in the RG population is plotted closer to the rest of the data.

After obtaining sequence information for each hotspot, we can compare the prevalence of various genic features in the hotspots with their degree of occurrence across the genome. More specifically, we look at the following features (many of which have been previously implicated in recombination in *Drosophila* or other organisms) in addition to GC/Exon content:

- CpG content - the fraction of nucleotides in a sequence that are part of a CpG or GpC dinucleotide
- Poly(A/T) content - the fraction of nucleotides that lie in a sequence of four or more consecutive Adenosine bases or a sequence of ≥ 4 consecutive Thymine bases
- Gene, Coding sequence (CDS), Transposable element (TE), transfer RNA (tRNA), and Transcription factor (TF) binding site contents - the fraction of each region annotated as part of the specified feature in the Flybase files (considering both DNA strands).

We find that exon, gene, and coding sequence prevalence in the hotspots tends to be moderately lower than the rest of the genome, while the presence of transposable elements is significantly diminished in hotspot regions (see Table 3). Also, there is nominal decrease in TF binding site content in the hotspots, although the amount fails to be significant due to the extreme variability of TF binding site content throughout the genome.

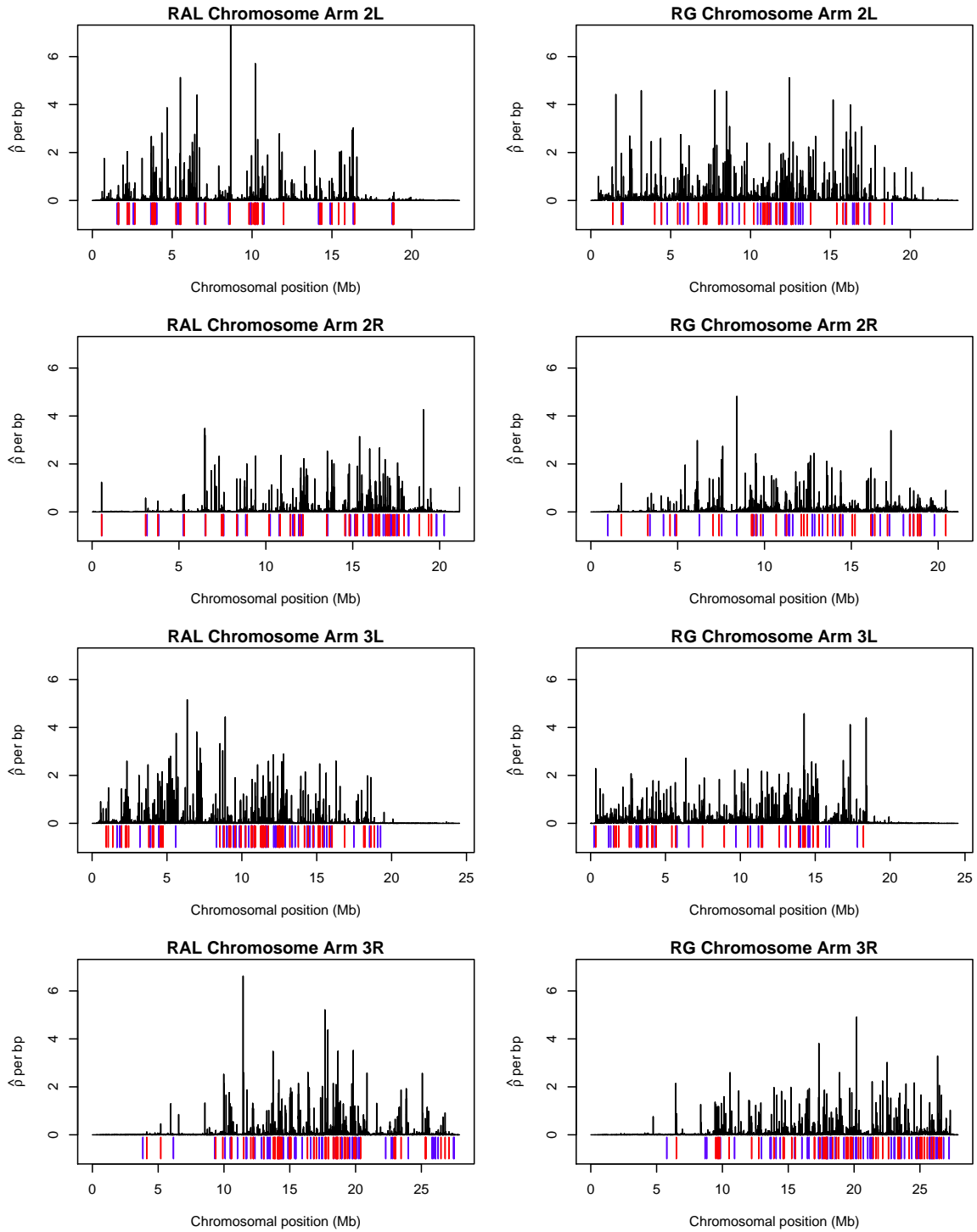


Figure 3: Recombination maps and hot/cold-spot locations. The black curves represent the recombination rates estimated by LDhelmet (using block penalty 10), while the red and blue bars denote the location of hotspot and coldspot regions, respectively.

genic feature	RAL hotspots	RAL genome	RAL P-value	RG hotspots	RG genome	RG P-value
GC content	0.419	0.428	0.97	0.426	0.429	0.97
CpG content	0.179	0.180	0.97	0.181	0.180	0.97
Poly(A/T) content	0.083	0.078	0.432	0.0757	0.0759	0.97
Exon content	0.118	0.159	0.067	0.117	0.159	0.067
Gene content	0.312	0.358	0.0968	0.306	0.358	0.067
CDS content	0.091	0.153	0.060	0.098	0.153	0.067
TE content	1.55e-3	0.032	< 0.009	4.56e-4	0.032	< 0.009
tRNA content	0.00	1.05e-4	0.970	1.35e-4	1.05e-4	0.970
TF binding site content	0.249	0.680	0.578	0.238	0.680	0.206

Table 3: Comparing genic features in hotspots with the rest of the genome. For each population: the hotspot columns contain the mean content of each feature across all hotspot sequences, the genome columns contain the genome-wide content of each feature, and the P-value columns contain false discovery rate (FDR) adjusted p -values for the significance of a two-sided test that the mean of the hotspot content values differs from the genome-wide content value (where FDR multiple testing correction is done via the Benjamini-Hochberg procedure [28] described in Supplementary Information). Due to non-normality of the genic feature content values, a t -test is not applicable in this situation, and we instead assess significance through a bootstrap procedure, in which 1000 bootstrapped datasets are generated under the null hypothesis of no difference by randomly sampling identically sized regions as each hotspot from the same chromosome arm. An estimate of the null distribution is then obtained by computing differences between the mean content value of each bootstrapped set of sequences and the genome-wide content value. Genic features for which this test is significant at the FDR-adjusted 0.1 level in both populations are highlighted in red (and because the bootstrap procedure is only repeated 1000 times, potentially minuscule p -values can only be identified as lying below the FDR-corrected level of 0.009).

IDENTIFYING DNA MOTIFS ENRICHED IN RECOMBINATION HOTSPOTS

Having established the location of hotspot regions and identified corresponding “coldspots” as representative sequences for the genomic regions lacking elevated recombination rates, we now search for DNA motifs which are overrepresented in the hotspots compared with the coldspots. A *motif* is defined as a short pattern (5-30 bp) of nucleotides, within which most bases are highly conserved at given positions while others can vary. It is important to note that because we must search through a vast configuration of DNA patterns to identify motifs of interest, multiple testing correction is imperative in assessing the significance of motif overrepresentation in hotspots, although being overly stringent in this regard results in overlooking an unacceptable number of potentially interesting motifs. Thus, while we test a multitude of motifs for significant overrepresentation in this section, multiple testing corrections are only performed with respect to the set of motifs found via the same source/method (presented in the same table) and we employ the less conservative Benjamini-Hochberg FDR approach for multiple testing correction [28].

We first examine previously characterized motifs which have been implicated in recombination as well as a libraries of experimentally-determined *D. melanogaster* transcription factor binding motifs available in the FlyFactorSurvey [29] and Jaspar CORE databases (<http://jaspar.genereg.net/> [30]). The motifs previously implicated in recombination as well as those established as transcription factor binding regions via a bacterial one-hybrid method in the FlyFactorSurvey are presented as consensus sequences (meaning degeneracies in the pattern are represented by IUPAC multi-base ambiguity characters rather than being quantified) and we adopt a simple approach described in [31] to assess whether they are overrepresented in hotspots: For each motif, we count the number hotspot sequences containing the motif as well the number of coldspots in which it is found, and create a 2×2 contingency table out of these values together with the number of hotspots without a motif match

and the number of coldspots without a match. Fisher’s exact test (based on the hypergeometric distribution) can then be employed to assess the null hypothesis that the sequences containing the motif pattern are evenly distributed between hotspots and coldspots against the alternative that there is a higher likelihood of a region containing the motif to exhibit high enough recombination rate to be identified as a hotspot rather than being selected as a coldspot. Furthermore, because a large number of motif copies in a region often strengthens their functional role (such as increasing the binding affinity of various molecules like transcription factors [31]), we also count the number of matches of each motif in the hotspots and in the coldspots. Again, we create a 2×2 contingency table by also counting the number of k -mers (where k is the motif length) which do not match the motif in the hotspots and in the coldspots, and we test the null hypothesis that the motif matches are evenly distributed between hotspots and coldspots again using a one-sided Fisher’s exact test. Note that all counts of motif occurrences/sequences containing motifs also take the reverse complement strand into consideration since features on the opposite strand are just as likely to be responsible for the peak in our recombination maps at the hotspots. We also considered counting patterns that are similar to the motif (with at most 1 or 2 differing nucleotides) as matches, but this produced overly large counts (with little difference between hot/cold spots) due to the brevity of the transcription factor binding motifs under consideration.

EXAMINING PREVIOUSLY IDENTIFIED MOTIFS

We investigate a number of previously characterized motifs, some of which Comeron et al. also examined for overrepresentation in the vicinity of estimated *D. melanogaster* cross-over sites rather than in recombination hotspots [20]. Since we simply wish to see whether our findings agree with those of previous studies, we do not employ multiple testing correction in our creation of Table 4. From the Table, we see that the human *PRDM9*-motif CCNCCNTNNCCNC is slightly enriched in the hotspots, although the amount is far from significant. Comeron et al. found significant enrichment of a shorter version of the motif, CCTCCCT, in the vicinity of estimated recombination events [20], and we also find that this motif is overrepresented in our hotspots, although only at the 0.05 significance level in the RAL population when considering the number of hotspots containing the motif vs. coldspots. We do not detect overrepresentation of TGACGT, a core hexamer motif in yeast which is bound by transcription factors (*Atf1* and *Pcr1*) that produce recombination-driving histone modifications in DNA leading to the formation of recombination hotspots in yeast where this motif is found [32]. Studying recombination-inducing motifs in *D. pseudoobscura*, a close relative of our *D. melanogaster*, Heil and Noor found that the motifs AATAAA and CTGCTG are weakly negatively associated with recombination on broad scales, while AAATTT and ACAAAT are weakly positively associated with recombination at the super-fine scale [15]. Our findings indicated that only AATAAA is slightly underrepresented in hotspots, while CTGCTG, AAATTT, and ACAAAT are fairly evenly distributed between hot and cold spots. Although Comeron et al. did not find support that GTGGAAA, a recently discovered motif significantly enriched near fifteen experimentally determined *Drosophila* crossover events [21], is overrepresented in the vicinity of the cross-over events they identified throughout the genome, we do find that this motif is nominally overrepresented in our hotspots. Finally, the least degenerate recombination-associated motif proposed by Comeron et al. is the poly (A) tail: AAAAAAAAAAAAAA, which we also find is significantly overrepresented in our hotspots identified from the RAL recombination map.

Motif consensus	Pop data	% Hotspot sequences	Seq ratio	Seq p-val	Hotspot motif #	Coldspot motif #	Motif ratio	Motif p-val
	RAL	7.4	0.93	0.65	15	15	1.0	0.57
CCNCCNT- -NNCCNC	Both	10.5	1.13	0.351	44	36	1.22	0.217
	RG	14.3	1.3	0.247	29	21	1.40	0.16
CCTCCCT	RAL	15.0	1.65	0.055	30	22	1.4	0.166
	Both	14.1	1.30	0.123	52	48	1.08	0.382
	RG	13.0	1.0	0.567	22	26	0.85	0.765
TGACGT	RAL	32.6	1.20	0.155	87	68	1.28	0.074
	Both	32.3	0.94	0.742	161	181	0.89	0.872
	RG	31.8	0.74	0.983	74	113	0.65	0.99
AATAAA	RAL	97.3	0.99	0.75	920	924	0.99	0.547
	Both	96	0.98	0.909	1779	1830	0.97	0.808
	RG	95	0.97	0.93	859	906	0.95	0.875
CTGCTG	RAL	79.7	1.04	0.266	403	406	0.99	0.556
	Both	81.8	1.04	0.193	856	845	1.01	0.404
	RG	47.4	1.03	0.324	453	439	1.03	0.331
CTCTCT	RAL	27	0.78	0.954	79	126	0.63	0.999
	Both	46.3	1.03	0.379	319	406	0.79	0.999
	RG	37	1.27	0.091	98	84	1.17	0.167
AAATTT	RAL	81	1.02	0.397	362	357	1.014	0.441
	Both	82	1.03	0.221	700	707	0.99	0.585
	RG	82	1.05	0.239	338	350	0.966	0.690
ACAAAT	RAL	81.3	1.02	0.397	724	714	1.01	0.406
	Both	90.9	1.02	0.304	512	449	1.14	0.245
	RG	92.2	1.03	0.276	517	548	0.94	0.838
GTGGAAA	RAL	36.9	1.17	0.163	93	85	1.09	0.300
	Both	37.2	1.04	0.375	176	185	1.95	0.701
	RG	37.7	0.92	0.758	83	100	0.73	0.909
AAAAAAA- -AAAAA	RAL	10.2	1.46	0.178	58	34	1.71	0.008
	Both	5.87	1.33	0.244	59	43	1.37	0.069
	RG	0.65	0.5	0.876	1	9	0.11	0.999

Table 4: Investigating hotspot overrepresentation of motifs implicated in recombination by previous studies. The Pop data column indicates whether we are examining the motif's presence in the hotspots/coldspots of the RAL or RG population (or both sets together), the % Hotspot sequences column contains the percentage of hotspot regions in which the motif is found, the Seq ratio column gives the ratio of number of hotspot sequences containing the motif to number of coldspot sequences containing the motif, the Seq p-val column contains p -values for a one-sided Fisher's exact test on the counts of hotspot/coldspot sequences containing the motif (as previously described), the Hotspot/Coldspot motif # columns contain the occurrences of matches to the motif in the hotspots/coldspots, the Motif ratio column gives the ratio of the number of motif occurrences in the hotspots to the number of occurrences in the coldspots, and the Motif p-val column contains the p -values of a one-sided Fisher's exact test on the counts of motif occurrences in hotspot/coldspot sequences (as previously described). p -values under the 0.1 unadjusted significance level are highlighted in red.

TRANSCRIPTION FACTOR BINDING MOTIFS

Applying the same counting methods to the 105,379 TF-binding consensus sequence motifs available in the FlyFactorSurvey database [29], we find that many of the motifs, whose p -values in both Fisher’s exact tests of motif-occurrence counts and number of motif-containing hotspot sequences are highly significant, have been identified as potential binding sites of the Antennapedia (*Antp*), caudal (*Cad*), aristaless (*Al*), or homeobrain (*Hbn*) transcription factors (see Table 5 and note that due to the extremely high number of binding sites characterized by the FlyFactorSurvey, any multiple-testing procedure for correcting the significance levels of individual motifs in this analysis adjusts all p -values to 1, and Table 5 therefore contains uncorrected individual p -values to avoid this loss of information). Investigating the function of these proteins in Flybase (flybase.org [33]), a comprehensive resource on *D. melanogaster* genes, we find that these transcription factors are all annotated as homeobox domains, meaning they have all been implicated in morphogenesis (the regulation of anatomical development) in *D. melanogaster*. We also examine each of these transcription factors in STRING, a database of known and predicted protein interactions which includes both direct (physical) and indirect (functional) associations, many of which are inferred by applying text-mining methods to the scientific literature [34]. Through text-mining, STRING identifies one of the predicted functional partners of aristaless to be mutagen-sensitive (*mus304*), a DNA damage checkpoint protein required for chromosome break repair which is implicated in reciprocal meiotic recombination in its biological process annotation. Furthermore, STRING’s text-mining method infers that the homeobrain protein has the following recombination-related functional partners (see Figure 4):

- Meiotic recombination 11 (*mre11*): involved in double-strand break repair via break-induced replication and non homologous end joining, and also implicated in reciprocal meiotic recombination
- Crossover suppressor on 2 of Manheim (*c(2)M*): has unknown function, but there is experimental evidence that it is involved in meiotic recombination
- N-acetyltransferase *eco* (*eco*): required for the establishment of sister chromatid cohesion, which is in turn required for homologous recombination
- Nipped-B protein (*Nipped-B*): plays a structural role in chromatin and supports sister chromatid cohesion, likely by interacting with the cohesion complex

Jaspar is another database of transcription-factor binding motifs in which each length- k motif is represented as a position weight matrix (PWM: $4 \times k$ matrix representing the distribution over the four possible bases at each given nucleotide position in a DNA motif) established through lab experimentation [30]. While we could collapse these PWMs into consensus sequences, this quantitative to qualitative transformation is accompanied by an undesired loss of information regarding the motif pattern. Frith et al. have proposed a PWM-based motif-finding method called Clover which scores the degree of the motif’s presence in a sequence via the average likelihood over all k -mers in the sequence, which is easily computed by sliding the motif’s PWM [31] along the target sequence. Motivated by a thermodynamic model, Frith et al. suggest that this matrix score reflects the motif’s binding energy at each location and they demonstrate the superiority of Clover over contingency-based methods both in simulation studies and in recovering well-known motifs from various sets of sequence data. In our analysis, we use the implementation of the Clover algorithm available in the PWMEnrich *R*/Bioconductor package [35] and we also employ another PWM-based significant-motif-identification algorithm implemented in the package, which is also based on binding affinity and uses a lognormal approximation to assess significance (but details are scant regarding this method because it has yet to be published [36]). However, we find that even

Motif consensus	Transcription factors	Pop data	% Hotspot sequences	Seq ratio	Seq p-val	Motif ratio	Motif p-val
ATTTATGA	AbdA, Antp, Cad,	Both	21.1	1.95	1.7e-4	2.13	1.0e-5
	Ubx, C15, HGTX,	RAL	19.8	1.95	6.6e-3	2.38	3.8e-4
	CG32105, CG34031, vvl, CG12361	RG	22.7	1.944	7.6e-3	1.92	5.7e-3
GCTTAATTN	Al	Both	18.2	2.21	8.2e-5	2.12	1.4e-4
		RAL	16.0	2.00	0.013	1.76	3.9e-2
		RG	20.8	2.46	1.7e-3	2.47	7.7e-4
AATTAATTN	Al	Both	28.4	1.67	2.5e-4	1.81	4.8e-9
		RAL	27.3	1.70	5.9e-3	1.88	1.1e-5
		RG	29.9	1.642	0.011	1.75	7.5e-5
CTTACTTA	Antp, Dfd, Eve, Zen, Hbn	Both	11.1	2.53	7.2e-4	3.13	2.9e-5
		RAL	11.8	2.75	6.1e-3	3.63	3.8e-4
		RG	10.4	2.29	0.040	2.57	0.022
TTAAATGC	Antp, Dfd	Both	25.8	1.66	6.3e-4	1.69	4.1e-4
		RAL	21.9	1.46	0.055	1.47	0.050
		RG	30.5	1.88	2.2e-3	1.94	1.6e-3
GCAAATTA	Cad, Tin, Hbn, Odsh, CG4328, CG33980	Both	25.5	1.89	5.0e-5	1.80	1.3e-4
		RAL	21.9	1.64	0.021	1.53	0.042
		RG	29.9	2.19	4.2e-4	2.07	5.5e-4
GCCAATTT	Hbn	Both	19.0	1.97	3.3e-4	2.11	1.2e-4
		RAL	18.2	1.62	0.039	1.57	0.059
		RG	20.1	2.58	1.4e-3	3.17	1.5e-4

Table 5: Transcription factor binding motifs from FlyFactorSurvey which are significantly enriched in hotspots at the unadjusted 0.001 level under both of the Fisher’s exact tests applied to the combined set of RAL/RG hotspot and coldspots. The Transcription factors column lists the *D. melanogaster* transcription factors which may bind to each motif according to experimental findings, the Pop data column indicates whether we are examining the motif’s presence in the hotspots/coldspots of the RAL or RG population (or both sets together), the % Hotspot sequences column contains the percentage of hotspot regions in which the motif is found, the Seq ratio column gives the ratio of number of hotspot sequences containing the motif to number of coldspot sequences containing the motif, the Seq p-val column contains p -values for a one-sided Fisher’s exact test on the counts of hotspot/coldspot sequences containing the motif (as previously described), the Hotspot/Coldspot motif # columns contain the occurrences of matches to the motif in the hotspots/coldspots, the Motif ratio column gives the ratio of the number of motif occurrences in the hotspots to the number of occurrences in the coldspots, and the Motif p-val column contains the p -values of a one-sided Fisher’s exact test on the counts of motif occurrences in hotspot/coldspot sequences (as previously described).

in the absence of multiple testing correction, none of the 125 transcription-factor binding motif PWMs in the Jaspas data are identified as significant by either method.

AB INITIO MOTIF DISCOVERY

Now, we turn our attention from previously characterized motifs to *ab initio* motif-discovery methods for identifying novel sequence patterns enriched in the hotspot regions, and we adopt three different approaches for this task. The first (and most widely used across a variety of motif-finding applications) is called MEME (Multiple Expectation Maximization for Motif Elicitation [37]), and this method uses the iterative approach of the expectation maximization (EM) algorithm to converge upon motifs that are unlikely to occur so repeatedly in the hotspot sequences by chance.

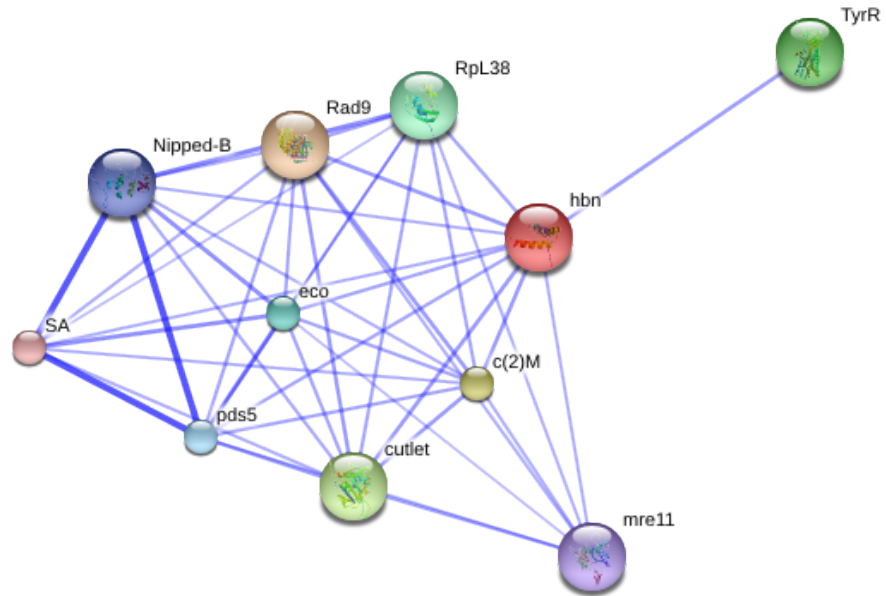


Figure 4: Predicted functional partners of the homeobrain (*hbn*) transcription factor determined by STRING’s text-mining method, which searches a large body of scientific texts including SGD, OMIM, FlyBase, and PubMed [34]. The thicker lines represent stronger inferred associations between proteins.

The hidden states in the EM procedure are the starting locations of the motif in each sequence, and at each iteration of the algorithm, MEME computes the likelihood of a given motif’s frequency in the hotspots using a Markov model on random background sequences. In our analysis, we run MEME with a first order background Markov model containing transition probabilities estimated from the observed frequencies of all nucleotides and dinucleotides across the entire genome of the sample from which the hotspot sequences are obtained. Once the EM procedure has converged to a motif whose frequency in hotspots has the lowest likelihood under the background model, it is masked and the algorithm re-initializes from a new starting point in the motif space, and we search for the top 10 most significant motifs via repeated application of this method.

MEME has primarily been applied for the discovery of transcription factor binding motifs and is rarely used with input sequences that are a couple of kb in length [38]. Since in our case, the length of numerous hotspot sequences exceeds the kb range, we must deal with a significant increase in running time, and because the performance of MEME on long sequences has not been well studied, we also performed the aforementioned word count Fisher’s exact test on each output motif included in Table 6 to verify its significance. Note that MEME measures motif significance using a likelihood ratio test (in which the null distribution is specified by our background Markov model) and reports Expect (E)-values, which are estimates of the number of motifs with the given likelihood ratio or higher that one would find in a similarly sized random set of sequences produced under the background Markov model. The E -value may be regarded as a multiple-testing corrected p -value as it takes the number of candidate motifs searched over into consideration. Interestingly, many of motifs output by MEME (presented in Table 6) are very similar to motifs found by Comeron et al. who applied MEME to the sequences in the vicinity of inferred crossover locations (Figure 6 of [20]): For example, Comeron et al. include PWMs representing the consensus sequences $(C/T)A(C/T)A(C/T)A(C/T)A(C/T)A(C/T)A(C/T)A(C/T)A(C/T)A(C/T)A$, $(A/T)NATANATATANATATAT$, $NGCCAACGCCC(A/C)$ in their list of significant motifs, and Table 6 contains PWMs representing the consensus sequences $CACACACACACAC$, $ATATA(T/C)-ATATATA$, and $GCCACGCCCCAC$.

In addition to MEME, we also use the DREME approach to search for motifs overrepresented










Position weight matrix	E-value	Pop data	Sites
	3.2e-50	Both	> 50
	2.9e-39	Both	49
	8.0e-36	Both	> 50
	2.0e-34	Both	> 50
	1.2e-34	Both	> 50
	1.8e-8	Both	44
	2.9e-39	RG	> 50
	5.4e-52	RAL	> 50
	9.4e-31	RAL	> 50

Table 6: The top-scoring *ab initio* motifs found by MEME. The *E*-value measures the significance of the likelihood ratio of each motif under a first order background Markov Model with parameters estimated from the (di)nucleotide frequencies across the genome of the most thoroughly sequenced RAL and RG samples (from which the hotspots sequences were obtained). We run MEME on the hotspots from the RAL samples separately from the hotspots of the RG group, as well as on the combined hotspot data, and the Pop data column specifies which set of input hotspot sequences each motif was discovered in.

in the hotspot regions (where despite sharing authors and similar names, these two approaches are entirely different). Although tailored for ChIP-seq data, DREME (discriminative regular expression motif elicitation [39]) presents a simple and effective method for finding statistically significant, discriminative motifs between our hot/cold spot sequences: First, all k -mers (for small k) are exhaustively enumerated in both sets of sequences, and the significance of the relative enrichment of each motif is calculated using the previously described Fisher’s exact test applied to counts of hot/cold spot sequences containing the motif. Subsequently, the highly significant discriminative motifs are combined into IUPAC regular expressions to produce a degenerate motif with increased significance. Finally, all k -mers matching the most significant regular expression in the hotspots are integrated to create a position weight matrix for the most significant motif. Like MEME, DREME then masks this motif from all sequences and re-initializes to find other differentially enriched sequence patterns.

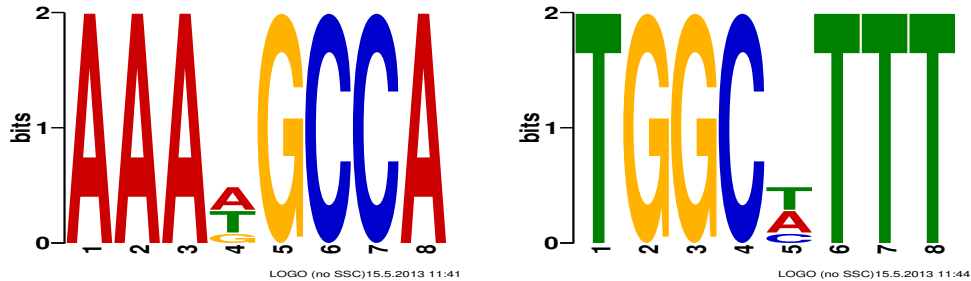


Figure 5: The only motif produced by DREME (on the left; its reverse complement is depicted on the right) which exceeds the E -value threshold of 0.05 when given all 341 RAL and RG hotspot/coldspot pairs as input. The height of each letter represents the information content of the corresponding nucleotide position in the motif.

Using the recommended threshold E -value of 0.05, DREME finds no significant motifs in the RG or RAL hot/coldspots when considered separately. Even when given all hot/coldspots from both sets of samples as input, DREME still only manages to find a single motif (with E -value = 0.0099, depicted in Figure 5) that meets the threshold. Collapsing the PWM into the consensus sequence AAADGCCA (where D is the IUPAC code for: A, T, or G) and counting motif occurrences in the hot/cold spots as well as the number of each of these types of regions which contains a copy of this consensus sequence, we find significant evidence that this motif is hotspot-enriched: in the RAL data, twice as many of the hotspots (26%) contain the motif as coldspots, and this results in $p = 0.0018$ for the test of the null hypothesis that the sequences containing this motif are evenly distributed between hot/cold spots vs. the alternative that these sequences are more likely to be identified as hotspots. Furthermore, this motif occurs 102 times in RAL hotspot sequences and only 87 times in RAL coldspots, a difference for which the one-sided test that the motif occurrences are distributed evenly has significance $p = 0.0012$. In the RG hot/coldspots, 1.5 times as many hotspot sequences contain the motif as coldspots ($p = 0.027$) and it occurs 1.35 as many times in hotspots as in coldspots ($p = 0.067$), where this set of four p -values for this motif has been FDR-adjusted.

The third *ab initio* discriminative motif-discovery method we employ is implemented in the R package MotifRG [40]. This regression-based approach works as follows: All k -mers are exhaustively enumerated (where k is a fixed, small value; we tried 5-10) and their discriminative power between hot/cold spots is measured using logistic regression. The top-ranking one of these k -mers is then iteratively refined by attempting to extend it one nucleotide at a time and also allowing IUPAC degenerate letters to be incorporated into its sequence. At each iteration, the slightly perturbed motif is evaluated against the current candidate motif and is adopted as the new candidate if it exhibits significantly greater discriminative power. Using the likelihood ratio test given by the logistic regression framework to assess statistical significance, the package also gives z-value statistics from permutation tests to reflect the discriminative power of each motif, and the motifs with the highest discriminative power are presented in Table 7. Note that the AAANGCCA motif output by DREME is present as the AAATGCCA consensus sequence in the candidate motifs identified by motifRG.

Because these three motif-discovery methods rely on different randomization techniques to extend short seed sequences into longer, more degenerate motifs with increased statistical overrepresentation, we find that the motifs output by one method can be very different than the patterns identified by the others. However, because DREME and motifRG search for motifs which discriminate well between the hotspots and coldspots, whereas MEME simply seeks patterns which









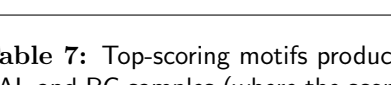
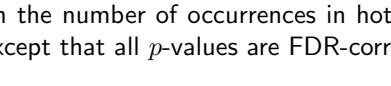
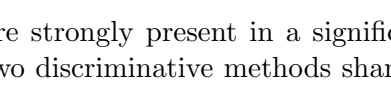
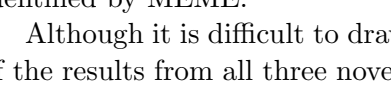
Position weight matrix	MotifRG Score	Pop data	% Hotspot sequences	Seq ratio	Seq p-val	Motif ratio	Motif p-val
	4.27	RAL	99.4	1.03	0.075	1.33	7.5e-9
		Both	98.8	1.02	0.103	1.31	5.8e-16
	3.89	RG	98.1	1.01	0.522	1.30	2.4e-8
		RAL	97.3	1.03	0.159	1.34	2.4e-8
	3.84	Both	97.9	1.02	0.077	1.26	1.1e-10
		RG	98.7	1.02	0.239	1.20	3.4e-4
	3.23	RAL	26.2	1.81	6.0e-3	1.74	0.013
		Both	27.0	1.92	1.2e-4	1.96	9.6e-5
	3.22	RG	27.9	2.05	3.4e-3	2.23	1.0e-3
		RAL	34.2	1.64	5.2e-3	1.85	1.2e-3
	3.18	Both	30.2	1.51	3.1e-3	1.63	1.0e-3
		RG	25.3	1.34	0.123	1.33	0.142
	3.22	RAL	21.9	2.05	4.8e-3	2.09	4.4e-3
		Both	24.6	1.87	3.7e-4	1.91	3.0e-4
	3.18	RG	27.9	1.72	0.014	1.77	0.014
		RAL	37.4	1.43	0.019	1.70	1.3e-3
	3.18	Both	39.0	1.37	4.8e-3	1.53	3.2e-4
		RG	40.9	1.31	0.061	1.38	0.033
	3.06	RAL	16.0	2.14	0.012	2.19	9.4e-3
		Both	19.1	1.91	1.4e-3	1.90	1.6e-3
	2.94	RG	22.7	1.75	0.025	1.70	0.036
		RAL	34.8	1.63	5.2e-3	1.69	1.1e-5
	2.94	Both	24.3	1.36	0.033	1.44	4.8e-4
		RG	11.7	0.86	0.769	0.766	0.906

Table 7: Top-scoring motifs produced by MotifRG when run on the combined set of hot/cold spots from both RAL and RG samples (where the scores in the second column reflect the discriminative power of the motifs based on the number of occurrences in hotspots vs. coldspots). The other columns are defined as in Tables 4 and 5, except that all p -values are FDR-corrected.

are strongly present in a significant fraction of the hotspot sequences, the motifs output by the two discriminative methods share much more resemblance with one another than with the motifs identified by MEME.

Although it is difficult to draw inferences from *ab initio* motif discovery, we can note that many of the results from all three novel motif discovery methods contain the common 4-mer GCCA core pattern. This motif has appeared in the literature in a variety of settings, but most prominently, Kwong et al. discovered that motifs comprising of a core GCCAT sequence with a less pronounced tail of four Ts attract certain Polycomb-group (*PcG*) genes, which build the Polycomb Repressive Complexes *Pcr1* and *Pcr2* needed for gene silencing in *Drosophila* [41]. Interestingly, *Pcr1* has been found to be instrumental in the formation of recombination hotspots in yeast [9], and possible recombination-related roles of the Polycomb group certainly warrant further investigation based on how extensively this core motif appears in our hotspots. Having identified a multitude of candidate recombination-inducing motifs which are enriched in the hotspot regions, we now investigate which of these motifs (as well as the other genic features overrepresented in the hotspots) are correlated with recombination rates across the genome.

CORRELATING MOTIFS AND FEATURES WITH RECOMBINATION RATES

Numerous previous studies have already examined global correlations between recombination rates and features of the *D. melanogaster* genome together with population characteristics [1, 16, 20], and we instead direct our efforts toward identifying local associations between features and recombination rates throughout the genome, exploiting the fine-scale resolution of the map provided by Chan et al. [1]. Because even *PRDM9*, the recombination-inducing motif with the strongest signal observed in any organism thus far, only occurs in around 40% of human recombination hotspots, it is very unlikely that our candidate motifs exhibit global correlation with recombination rates, and we therefore focus on identifying motifs and features which are highly associated with local variation in the recombination map at intermittent regions of the *D. melanogaster* genome. Treating the each chromosome arm as a temporal rather than a spatial sequence, we search for local relationships via the method of *wavelet coherence analysis* (described in the Supplementary Information and more thoroughly in [42]), which has been very successfully applied to transform wildly-oscillating time-series data to a highly intuitive form in which transient relationships between different series become easy to discern. Briefly, wavelet coherence offers a smoothed, local measure of correlation between the continuous wavelet expansions of two different time series on the same domain (genic feature values along the DNA sequence of each chromosome arm in this case) [42]. By plotting wavelet coherence in the time-frequency domain, it is easy to identify areas in which the two series exhibit common power and consistent phase behavior, features which suggest a local relationship between the two signals.

Following Chan et al., who employed a number of wavelet methods in their analysis, we first bin our log-transformed recombination rate estimates into 250 bp windows to manipulate each of our recombination maps into a suitable time series format [1]. To obtain series of interesting genic features that can be analyzed against the RG/RAL recombination rate series, we calculate the following continuous-valued measures for each sequence window:

- Enrichment of one of our previously identified candidate motifs, which is computed as the Clover average likelihood score across all possible alignments of the motif’s PWM against the window (Frith et al. suggest this provides a good measure of its binding affinity) [31]. Note that this method can be applied for motifs represented as consensus sequences by simply assuming their PWM distribution places extremely high probability near 1 on each base in the consensus sequence (and correspondingly reduced probabilities for IUPAC ambiguous characters in the consensus, such as uniform $\frac{1}{4}$ probability over A,T,C,G for each base called as N).
- Distance from one end of the window to the nearest occurrence of a given candidate motif, where motif occurrence is assumed at all k -mers with Clover likelihood scores strictly above the threshold score of a k -mer which only differs from the motif at one non-degenerate base (and the distance is measured as 0 if there is such an occurrence within the window).
- Content of exons as well as transposable elements (TEs).
- Distance from one end of the window to the nearest gene and the distance to the nearest transposable element (0 if there is such a feature inside the window). While genic distances are technically discrete, most of the distances are on the order of thousands of base-pairs and this divergence from the continuity of series values assumed in the wavelet methods is thus indiscernible.

- Distance to the nearest H2A.Z and bulk nucleosomes from the window as well as bulk and H2A.Z nucleosome content within the window (where bulk nucleosomes are defined to be those containing any combination of H2A.Z and H2A). The locations of these two types of nucleosomes are taken from the high-resolution *Drosophila* nucleosome map produced by the Penn State Genome Cartography Project (<http://atlas.bx.psu.edu/> [43]), which was inferred through high-density tiling array experimentation. Nucleosomes, which consist of a compactly packaged DNA sequence wound around histone cores, enable eukaryotic organisms to carry meter long DNA sequences in their microscopic cells and it is widely acknowledged that the numerous chromatin modifications present in the sequence around these structures strongly influence local recombination rates [21]. Furthermore, Mavrich et al. have demonstrated that H2A.Z and bulk nucleosomes have dissimilar effects on neighboring chromatin structure and exhibit very different types of interactions with regulatory elements [43], and we thus investigate whether there are differing relationships between recombination rates and these two types of nucleosomes.

We utilize the wavelet coherence methods our own slightly-modified version the *biwavelet R* package (which in turn is based on the WTC Matlab package by A. Grinsted [42]) to compare our recombination rate “time series” with each of the above mentioned series of genic feature measurements [44]. Both our recombination rate “time series” and our genic feature series are first decomposed into continuous wavelet representations consisting of coefficients indexed by position (“time”) and scale which represent the variation in the input series at each position/scale. For every position and scale (we use period rather scale for the plots), the wavelet coherence is computed, which can be thought of as a squared correlation coefficient between the variation in the two series at the given position/scale [1], and we then examine where in the position-scale domain there are correlations between the local power of the series. To assess whether the observed coherence at a given position/scale is statistically significant, *biwavelet* generates an empirical null distribution of 1000 sets of coherence values using Monte Carlo simulation in which both series are generated by autoregressive processes of order 1 with underlying red noise spectra as described in [42]. Subsequently, a χ^2 -test is performed for the observed coherence values from each scale/time, in which the unknown expected amounts of coherence are substituted by the empirical means of the Monte Carlo simulated values.

By examining numerous wavelet coherence plots comparing series of motif enrichment value with the recombination rate series, we find that for our data, the discriminative MotifRG and DREME approaches tend to produce motifs with much larger regions of significant coherence in the time-frequency domain of each chromosome arm than the motifs output by MEME, despite the fact that MEME accounts for genome-wide background nucleotide pattern distribution (albeit via simplified first order Markov Model) while the discriminative methods simply compare candidate motifs in hotspots against the set of randomly selected coldspot sequences. For brevity, only plots we found interesting are included in this paper and the majority focus solely on Chromosome arm 2L of the samples from each population (as we find the degree of coherence between most genic features and recombination rates is fairly similar across all chromosome arms). Note that the period-labeling of the y -axis in Tables 6-11 is an approximation of the underlying Fourier period which corresponds to the oscillations in the wavelet, and there is a one-to-one relationship between period and scale [42].

Figure 6 does not suggest a clear difference between the role of the H2A.Z and bulk nucleosomes, but it is clear that at medium scales, there is a significant relationship between nucleosome occupancy and recombination rate variation in various regions along chromosome arm 2L. We find similar results in the other three chromosome arms as well as in the wavelet coherence plots be-

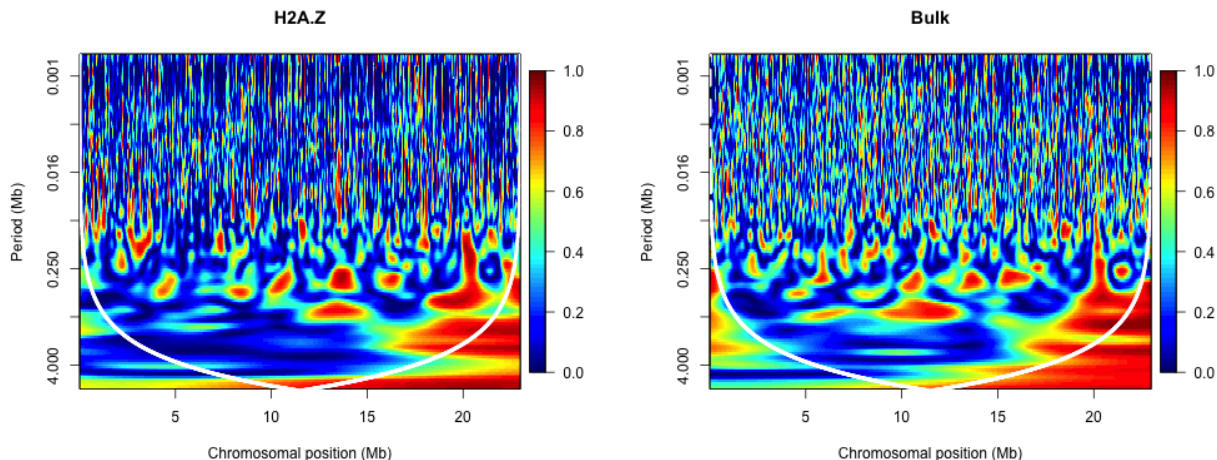


Figure 6: Wavelet coherence between recombination rates and Bulk/H2A.Z nucleosome content. Both plots depict the coherence between the recombination and one of the nucleosome maps for chromosome arm 2L of the RAL samples. Shown in white is the cone of influence, beyond which the wavelet transform is subject to significant distortions due to edge effects and thus the coherence values in this area of the time-frequency domain may be incorrect [42].

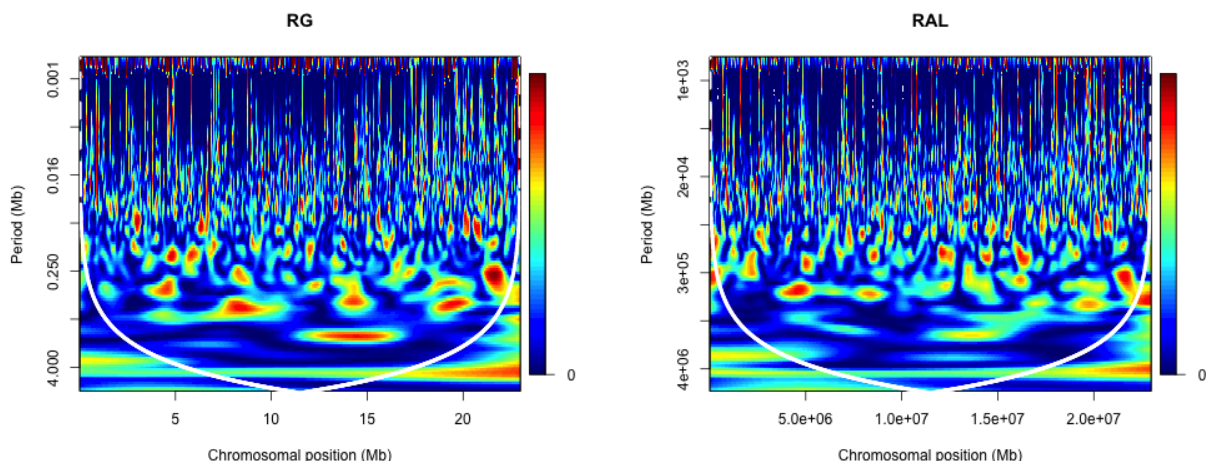


Figure 7: Wavelet coherence between recombination rates and distance to nearest transposable element in chromosome arm 2L of the RAL and RG samples. The smallest period used in the wavelet decomposition is 500 bp shown along the top of the plots, and the cone of influence is represented in white.

tween recombination rates and the series where distance to the nearest nucleosome is measured rather than nucleosome content. In fact, agreeing with the results of [45], our recombination rate estimates tend to be elevated in regions of low nucleosome occupancy which are farther from the nearest nucleosome.

Due to the limited number of transposable elements in the annotations in the *D. melanogaster* genome, investigating coherence between recombination rates and TE content (which is zero almost everywhere) did not produce interesting plots. However, when examining the wavelet coherence plot of recombination rate and distance to the nearest TE (Figure 7), we find that there is a very significant relationship between these quantities at extremely fine scales. Looking further

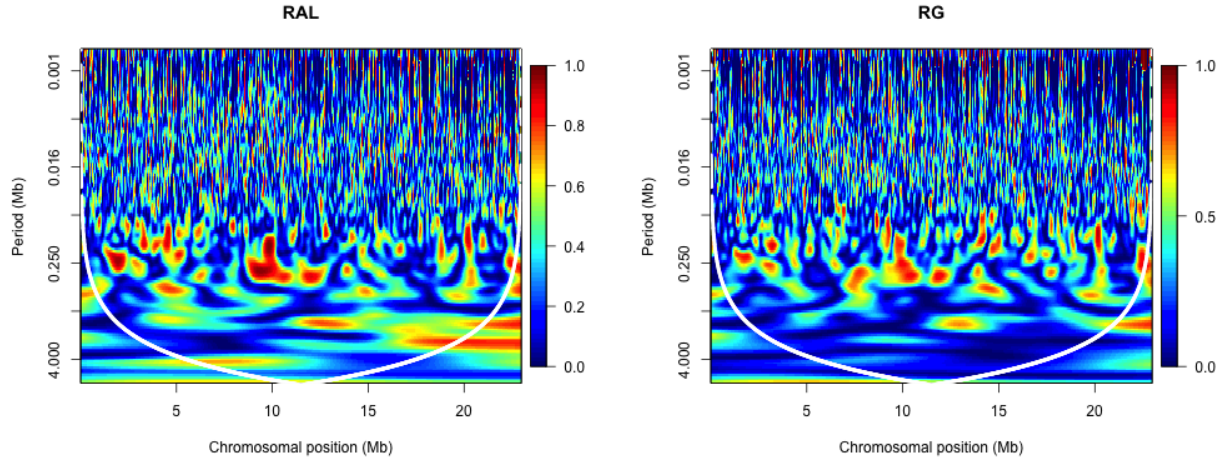


Figure 8: Wavelet coherence between recombination rates and distance to nearest gene annotated region in chromosome arm 2L of the RAL and RG samples. The cone of influence is shown in white.

into this relationship, we discover that regions closer to TEs exhibit suppressed recombination rates, which supports our earlier finding of TE underrepresentation in the recombination hotspots. While neither Exon/Gene content nor distance to nearest Exon exhibited significant coherence with recombination rates over any sizable region of the position-frequency domain, Figure 8 does show that at intermittent locations throughout arm 2L, the wavelet representation of distance to nearest gene does exhibit a clear relationship with the recombination rate wavelet transform, especially at the finest scales. We find that as in the vicinity of transposable elements, recombination rates in regions close to genes are more likely to be reduced.

We also examine the extent of coherence between recombination rates and each of our previously characterized motifs in the position-frequency domain to identify motifs which demonstrate the most significant relationship with recombination rates in local regions throughout the genomes of our RAL/RG samples. Interestingly, we find that the motif whose Clover scores exhibit the most significant coherence with local recombination rates in the wavelet domain (across all four chromosome arms in the samples from both populations), is the 8-mer GCCAATTT, characterized as a transcription factor binding site of the homeobrain (*Hbn*) gene in the FlyFactorSurvey database (see Table 5). This result is surprising because a number of the novel motifs identified through *ab initio* methods were far more significantly overrepresented in the hotspots compared with our randomly selected coldspot regions under the Fisher’s exact test comparison. Nevertheless, Figures 9,10, and especially 11 depict a strong recurring relationship (especially at the finer scales) exhibited between this motif and the local recombination map, both when its presence is measured by the Clover binding-affinity based likelihood ratio score or by distance to the nearest occurrence of the pattern. Similar to the discovery of the *PRDM9* motif, which Myers et al. found to be active in 40% of human recombination hotspots, we find the GCCAATTT consensus sequence present in 37% of our putative *D. melanogaster* recombination hotspots [10].

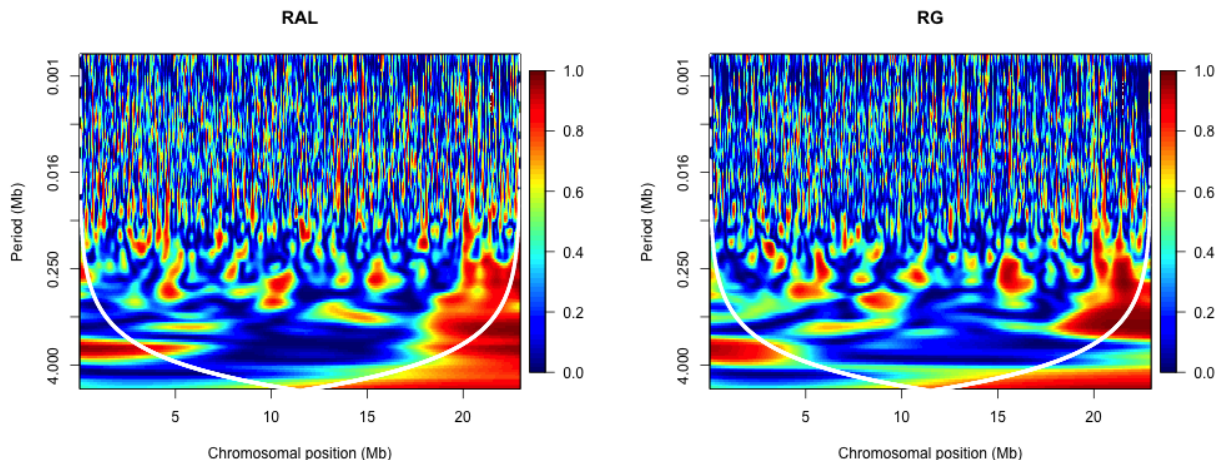


Figure 9: Wavelet coherence between recombination rates and Clover likelihood ratio scores of the GCCAATTT motif across chromosome arm 2L of the RAL and RG samples. The cone of influence is shown in white.

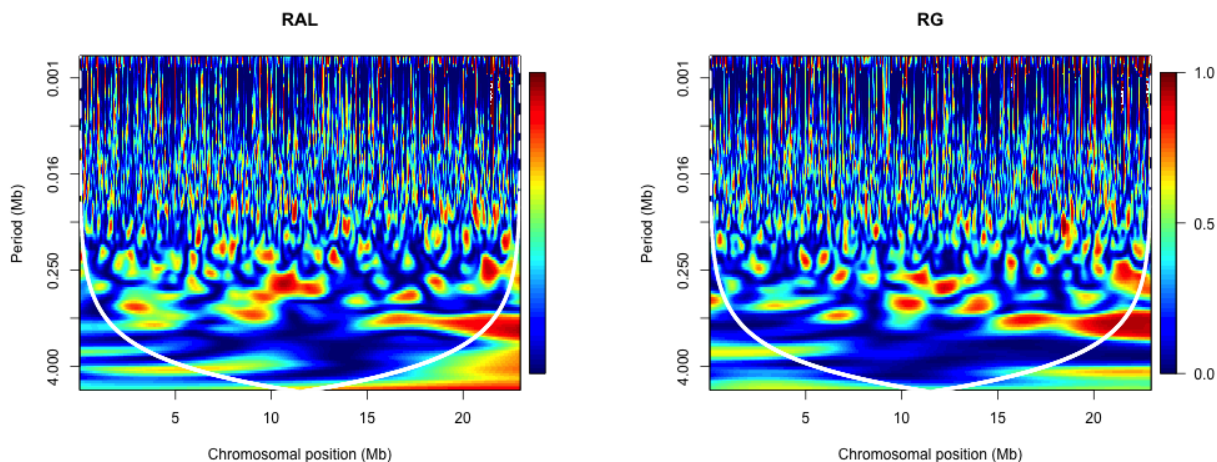


Figure 10: Wavelet coherence between recombination rates and distances to the nearest occurrence of an GCCAATTT 8-mer across chromosome arm 2L of the RAL and RG samples. The cone of influence is shown in white.

DISCUSSION

Our analysis has produced a number of intriguing findings regarding the relationship between recombination rates and features in the genome of *D. melanogaster*. We find that in accordance with the results in [46], transposable elements tend to either repress nearby recombination rates or be drawn to the vicinity recombination-repressed regions. In their study, Rizzon et al. suggest two differing explanations for the accumulation of TEs in the regions with lower recombination rate which are both based on the notion that selection is a confounding factor expected to be weaker in areas of reduced recombination and higher TE density. Their first hypothesis attributes the negative correlation between TE content and recombination rates to the fact that selection acts against deleterious mutations caused by TE insertions, and their other explanation is that selection may act against chromosomal rearrangements caused by ectopic recombination between

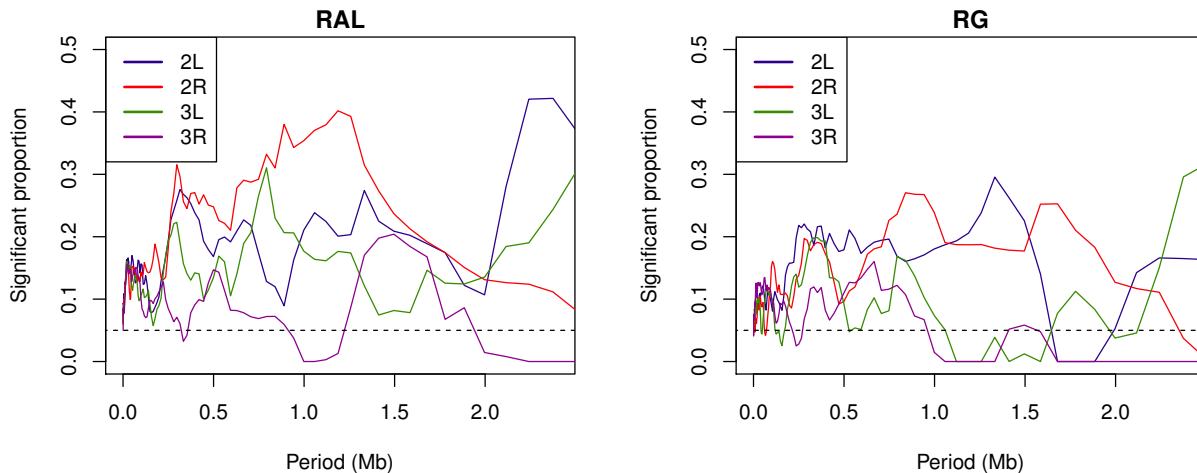


Figure 11: The significance of the wavelet coherence values between recombination rates and Clover scores of the GCCAATTT motif across the *D. melanogaster* genome in RAL/RG samples. For each chromosome arm and each wavelet period, the location on the corresponding curve denotes the proportion of sequence windows in our motif-score “time series” which exhibit statistically significant coherence with recombination rates at the 0.05 level.

TE copies. [46]

Our analysis strongly implicates the homeobrain transcription factor binding motif GCCAATTT as a recombination-inducing DNA sequence pattern, a finding which has not been suggested in the literature. However, it is well known that the specificity of a single homeodomain transcription factor is generally insufficient for solely recognizing its target genes and thus such proteins typically act in the promoter areas as part of a complex with other proteins. Therefore, it is certainly possible that other homeodomain proteins are also drawn to this motif and are the entities responsible for the recombination rate spikes inferred in the vicinity of this motif. Furthermore, we note that a significant fraction of the motifs we discovered via different approaches contain the core GCCA 4-mer, and it may be that an extension of this pattern similar to the GCCAATTT motif is the true feature behind the elevated rates in the vicinity of the motif we present. To further refine our motif search, we might follow the methods of Myers et al., in which degeneracy in the candidate motif is iteratively increased at all possible positions and extended until peak statistical significance is attained [10].

One shortcoming in our analysis is the fact that the high SNP density between the samples is neglected in our search for motifs. If there are in fact patterns in the *Drosophila* genome responsible for inducing recombination, the high degree of nucleotide variation in this organism reduces the likelihood that such motifs are preserved across the genomes of different flies and could explain the lack of common locations of extremely frequent recombination which are present in organisms with lower SNP densities such as mammals, plants, and yeast. One possible way to validate our motifs in the sequences of numerous samples would be to experimentally determine the locations of a few crossover events in numerous genomes, as was done by Miller et al. [21]. By determining which of the hotspot-enriched motifs are also overrepresented in the vicinity of single recombination events, one might be able to find a motif implicated in all *Drosophila* recombination, rather than being limited to the identification of patterns which potentially drive the mechanisms behind recombination hotspots.

The existence of a common mechanism which initiates the recombination process remains an open question in a large number of higher eukaryotes [14], and understanding the phenomena behind the non-uniformity in *Drosophila* recombination rates will be an important milestone in the quest for an answer. Comeron et al. summarize their *D. melanogaster* motif search by stating that while human and mice hotspots are associated with highly delimited genomic regions and a restricted number of DNA motifs, their data suggests a softer, more probabilistic landscape with an excess of recombination events within larger regions and a large and heterogeneous population of motifs [20]. If, in fact, the recombination landscape in *Drosophila* is far more complex than previously presumed and a large number of transcription factors with interactions between them is responsible for the heterogeneity of recombination-associated motifs, new methods are needed to identify these relationships from recombination-rate estimates. Nevertheless, we have discovered a number of strong associations between motifs, genic features, and recombination rates, some of which may lead to an improved picture of *Drosophila* recombination, and these computational findings should therefore be experimentally investigated in the near future.

ACKNOWLEDGEMENTS

I am deeply grateful to Professor Song for introducing me to the exciting field of computational biology and being a supportive mentor throughout the course of my undergraduate career. Thanks also to Andrew Chan for providing assistance with the recombination maps and Flybase annotations, as well as the administrators of the Berkeley Statistical Computing Facility for maintaining the computer grid on which the analysis was run.

REFERENCES

- [1] Chan AH, Jenkins PA, Song YS (2012) Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genetics* 8: e1003090.
- [2] Lichten M, Goldman AS (1995) Meiotic recombination hotspots. *Annual Review of Genetics* 29: 423–44.
- [3] Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics* 29: 217–22.
- [4] Edlinger B, Schlögelhofer P (2011) Have a break: determinants of meiotic DNA double strand break (DSB) formation and processing in plants. *Journal of Experimental Botany* 62: 1545–63.
- [5] Keeney S, Kleckner N (1996) Communication between homologous chromosomes: genetic alterations at a nuclease-hypersensitive site can alter mitotic chromatin structure at that site both in cis and in trans. *Genes to Cells* 1: 475–489.
- [6] Ohta K, Shibata T, Nicolas A (1994) Changes in chromatin structure at recombination initiation sites during yeast meiosis. *The EMBO Journal* 13: 5754–63.
- [7] White MA, Dominska M, Petes TD (1993) Transcription factors are required for the meiotic recombination hotspot at the HIS4 locus in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences* 90: 6621–6625.
- [8] Petes TD (2001) Meiotic recombination hot spots and cold spots. *Nature Reviews Genetics* 2: 360–9.
- [9] Gao J, Davidson MK, Wahls WP (2008) Distinct regions of ATF/CREB proteins Atf1 and Pcr1 control recombination hotspot ade6-M26 and the osmotic stress response. *Nucleic Acids Research* 36: 2838–51.
- [10] Myers S, Freeman C, Auton A, Donnelly P, McVean G (2008) A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics* 40: 1124–9.
- [11] Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, et al. (2010) Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* 327: 876–9.
- [12] McVean G, Myers S (2010) PRDM9 marks the spot. *Nature Genetics* 42: 821–2.
- [13] Baudat F, Buard J, Grey C, al E (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327: 836–40.
- [14] Goodstadt L, Ponting C (2011) Is the control of recombination conserved among diverse eukaryotes? *Heredity* 106: 710–11.
- [15] Heil CSS, Noor AF (2012) Zinc Finger Binding Motifs Do Not Explain Recombination Rate Variation within or between Species of *Drosophila*. *PLoS One* 7: e45055.
- [16] McKim KS, Jang JK, Manheim EA (2002) Meiotic recombination and chromosome segregation in *Drosophila* females. *Annual Review of Genetics* 36: 205–232.

- [17] 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: 321–4.
- [18] Singh ND, Aquadro CF, Clark AG (2009) Estimation of fine-scale recombination intensity variation in the white-echinus interval of *D. melanogaster*. *Journal of Molecular Evolution* 69: 42–53.
- [19] Cirulli ET, Kliman RM, Noor MAF (2007) Fine-scale crossover rate heterogeneity in *Drosophila pseudoobscura*. *Journal of Molecular Evolution* 64: 129–135.
- [20] Comeron JM, Ratnappan R, Bailin S (2012) The Many Landscapes of Recombination in *Drosophila melanogaster*. *PLoS Genetics* 8: e1002905.
- [21] Miller DE, Takeo S, Nandanan K, Paulson A, Gogol MM, et al. (2012) A Whole-Chromosome Analysis of Meiotic Recombination in *Drosophila melanogaster*. *G3: Genes-Genomes-Genetics* 2: 249–60.
- [22] McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, et al. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* 304: 581–4.
- [23] Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider D, et al. (2012) Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* 192: 533–98.
- [24] McQuilton P, Pierre SES, Thurmond J (2012) FlyBase 101the basics of navigating FlyBase. *Nucleic Acids Research* 40: D706–14.
- [25] Bergero R, Charlesworth D (2009) The evolution of restricted recombination in sex chromosomes. *Trends Ecol Evol* 24: 94–102.
- [26] Jensen MA, Charlesworth B, Kreitman M (2002) Patterns of genetic variation at a chromosome 4 locus of *Drosophila melanogaster* and *D. simulans*. *Genetics* 160: 493–507.
- [27] Fearnhead P (2006) SequenceLDhot: detecting recombination hotspots. *Bioinformatics* 22: 3061–66.
- [28] Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J R Statistic Soc* 57: 289–300.
- [29] Zhu LJ, Christensen RG, Kazemian M, Hull CJ, Enuameh MS, et al. (2011) FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Research* 39: D111–7.
- [30] Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, et al. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Research* 36: D102–6.
- [31] Frith MC, Fu Y, Yu L, Chen JF, Hansen U, et al. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Research* 32: 1372–81.
- [32] Fox ME, Yamada T, Ohta K, Smith GR (2000) A Family of cAMP-Response-Element-Related DNA Sequences With Meiotic Recombination Hotspot Activity in *Schizosaccharomyces pombe*. *Genetics* 156: 59–68.

- [33] Marygold SJ, Leyland PC, Seal RL, Goodman JL, Thurmond JR, et al. (2013) FlyBase: improvements to the bibliography. *Nucleic Acids Research* 41: D751–D757.
- [34] Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research* 39: D561–8.
- [35] Stojnic R (2012) PWMEnrich: PWM enrichment analysis. *R package (Version 1.0.2)* .
- [36] Stojnic R, Adryan B (2012) Identification of functional dna motifs using a binding affinity lognormal background distribution. *Submitted* .
- [37] Bailey TL, Boden M, Buske Fa, Frith M, Grant CE, et al. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* 37: W202–8.
- [38] Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology* 23: 137–44.
- [39] Bailey TL (2011) DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 27: 1653–9.
- [40] Zizhen Y (2012) motifRG: A package for discriminative motif discovery, designed for high throughput sequencing dataset. *R package (Version 2.12)* .
- [41] Kwong C, Adryan B, Bell I, Al E (2008) Stability and Dynamics of Polycomb Target Sites in *Drosophila* Development. *PloS Genetics* 4: e1000178.
- [42] Grinsted A, Moore JC, Jevrejeva S (2004) Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics* 11: 561–66.
- [43] Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, et al. (2008) Nucleosome organization in the *Drosophila* genome. *Nature* 453: 358–62.
- [44] Gouhier T (2013) biwavelet: Conduct univariate and bivariate wavelet analyses. *R package (Version 0.14)* .
- [45] Yin J (2010) Computational Methods for Meiotic Recombination Inference. *Technical Report No UCB/EECS-2010-169* .
- [46] Rizzon C, Marais G, Gouy M, Biemont C (2002) Recombination Rate and the Distribution of Transposable Elements in the *Drosophila melanogaster* Genome. *Genome Research* 12: 400–407.
- [47] Torrence C, Webster P (1998) A practical guide to wavelet analysis. *Bull Am Meteorol Soc* 79: 61–78.

SUPPLEMENTARY INFORMATION

COMPARING BLOCK PENALTY 10 AND BLOCK PENALTY 50 MAPS

To investigate whether the LDhelmet block penalty 10 results are reasonable, we smooth the resulting estimated recombination rates (using a sliding “moving average” window of 5 kb) to reduce the increase in variation associated with the lower block penalty and then compare the smoothed values with the block penalty 50 estimates. From Figure S1, it is evident that after smoothing has reduced all the excess variation associated with the lower block penalty, the peaks of the maps under different block penalties are virtually superimposed in the depicted region, and we found very similar results in a multitude of different-sized windows across the genome in both populations. Thus, while lowering the block penalty results in much more variability in the estimated recombination rates enabling more powerful detection of hotspots, the reduction in block penalty is not accompanied by a growth in the number of spurious recombination spikes that do not align with the recombination peaks in the BP-50 recombination map.

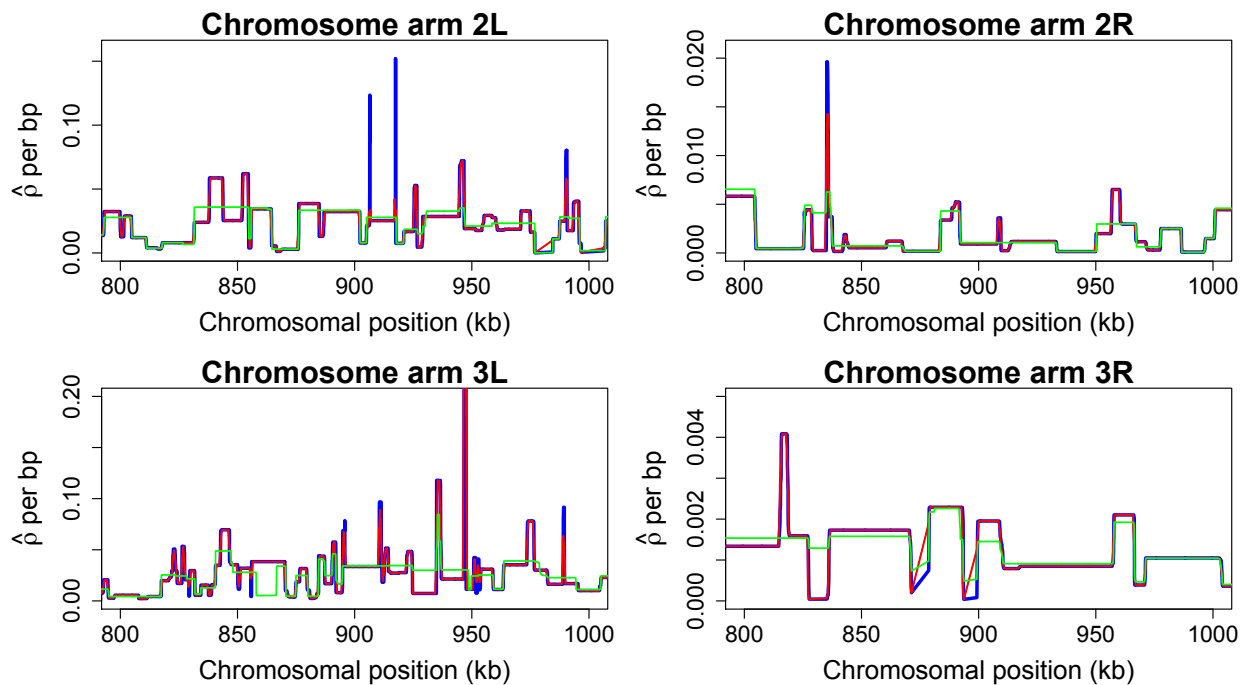


Figure S1: Comparing LDhelmet recombination maps produced using block penalties 50 and 10 for each chromosome arm of the RAL samples. The genomic region depicted here is the last 200 kb of the first megabase of each chromosome arm, which is the first location where the recombination rates begin to exhibit marked variability across all four arms. The blue curve (plotted underneath the others) denotes the block penalty 10 estimated recombination map, the red curve denotes the original BP-50 map, and the green curve represents a 5 kb sliding average smoothed version of the map estimated under block penalty 10.

FALSE DISCOVERY RATE MULTIPLE-TESTING CORRECTION

In this work, we account for multiple-testing by employing the Benjamini-Hochberg (BH) procedure, which is suitable for controlling the false discovery rate (FDR) across a wide range of problems [28]. Unlike multiple-comparison approaches that control the familywise error rate (FWER: the probability of one or more false positives) such as the Bonferroni procedure, the BH method controls the expected proportion of false null hypothesis rejections in the family of tests and is often adopted to correct large families of tests, as it has been noted that FWER procedures tend to be overly conservative in this scenario [28]. The BH method involves the following steps:

- Let q denote the number of statistical tests in consideration, let $p_{(1)}, p_{(2)}, \dots, p_{(q)}$ be the significance levels of the q comparisons (in increasing order), and let α be the desired combined significance level (generally 0.1, 0.05, or 0.001 in this work)
- Compute $j^* = \min\{j \in \{1, \dots, q\} : p_{(j)} \leq \frac{j \cdot \alpha}{q}\}$
- Declare all tests associated with $p_{(1)}, \dots, p_{(j^*)}$ significant and the fail to reject the null hypothesis for the tests associated with $p_{(j^*+1)}, \dots, p_{(q)}$

Benjamini and Hochberg argue that this method is an arguably more appropriate approach for identifying a few important effects from a multitude of comparisons than FWER approaches [28].

WAVELET COHERENCE ANALYSIS

Wavelets are simple zero-mean functions which are used to split a continuous time-signal into different scale components via the wavelet transform, in which the time-series is represented by a sum of wavelets from the same family with varying scaling parameters [42]. Each wavelet is defined by a mother function (describing its family) and a scaling function, which controls the coverage of wavelet spectrum. Note that while we use the term *scale* to refer to the width of a wavelet and the term *period* to describe the approximate Fourier period corresponding to the wavelet's oscillations, there is a one-to-one relationship between these quantities. When a given wavelet's scale factors are low, it offers a detailed representation of the time-signal, but this is accompanied by the trade-off of reducing the amount of the time domain covered by this wavelet. As Chan et al. mention, one distinct advantage of the wavelet approach is that it circumvents specification of window sizes, which would otherwise be needed to split the recombination maps into local regions for methods such as regression [1]. To compute the wavelet coherence between two time series, a continuous wavelet transform is first applied to each dataset, in which a chosen wavelet is applied as a band-pass filter to the series while varying its scale so that it is stretched in time.

The continuous wavelet transform employed in our analysis (after the recombination rates are put into a time-series format by binning their values into 250 bp log-transformed windows) uses the Morlet mother wavelet, which is widely adopted due to its simplicity and descriptive time and frequency localization [42]. Furthermore, the Morlet wavelet has several smooth oscillations and a well-defined period, which is a good approximation of the Fourier period in the time-signal, unlike the period of many other wavelets. As cautioned in [1], data from distant regions influences the wavelet transform at each position (time/nucleotide location) proportionally to the scaling factor, and thus sizeable regions of the time-frequency domain in an area known as the *cone of influence* are distorted by undesired effects (especially at larger scales) as a result of the inherent discontinuity at the edge of the region for which we have data. More precisely, we follow [1, 42] defining the cone of influence as the region of the time-frequency domain where wavelet power for an edge discontinuity falls to e^{-2} of the value at the edge, and we shift our focus away from the possibly-distorted wavelet results within this region.

Subsequently, the cross wavelet transform is computed for each time/frequency by multiplying the wavelet coefficient of one series with the conjugate wavelet coefficient of the other, and this method can be used to reveal areas with high common power. Given a pair of time series X and Y , we define $W_X(n, s)$ to be the continuous wavelet transform representation for X at time point n and scale value s (and $W_Y(n, s)$ is similarly defined for Y). The wavelet coherence between the two wavelet transforms is given by:

$$R^2(n, s) = \frac{\left| S(s^{-1}W_X(n, s) \cdot \overline{W_Y(n, s)}) \right|}{S\left(s^{-1}|W_X(n, s)|^2\right) \cdot S\left(s^{-1}|W_Y(n, s)|^2\right)}$$

where S is a smoothing operator and $\overline{\cdot}$ denotes complex conjugation. Thus, the wavelet coherence closely resembles a smoothed “correlation” coefficient [42]. In our analysis, the smoothing is done both along scale and time axes, using the smoothing operator for the Morlet wavelet suggested by Torrence and Webster [47].