

The fine-scale recombination rate variation and associations with genomic features in a butterfly

Aleix Palahí i Torres^{1,*}, Lars Höök¹, Karin Näsvall¹, Daria Shipilina¹, Christer Wiklund², Roger Vila³, Peter Prusscher¹ and Niclas Backström¹

¹Evolutionary Biology Program, Department of Ecology and Genetics (IEG), Uppsala University, Sweden. Norbyvägen 18d, SE-752 36, Uppsala, Sweden

²Department of Zoology: Division of Ecology, Stockholm University, Svante Arrhenius väg 18B, SE-106 91 Stockholm, Sweden

³Butterfly Diversity and Evolution Lab, Institut de Biologia Evolutiva (CSIC-UPF), Barcelona, Spain

*Corresponding author: aleix.palahi@ebc.uu.se

Abstract

Genetic recombination is a key molecular mechanism that has profound implications on both micro- and macro-evolutionary processes. However, the determinants of recombination rate variation in holocentric organisms are poorly understood, in particular in Lepidoptera (moths and butterflies). The wood white butterfly (*Leptidea sinapis*) shows considerable intraspecific variation in chromosome numbers and is a suitable system for studying regional recombination rate variation and its potential molecular underpinnings. Here, we developed a large whole-genome resequencing data set from a population of wood whites to obtain high-resolution recombination maps using linkage disequilibrium information. The analyses revealed that larger chromosomes had a bimodal recombination landscape, potentially due to interference between simultaneous chiasmata. The recombination rate was significantly lower in subtelomeric regions, with exceptions associated with segregating chromosome rearrangements, showing that fissions and fusions can have considerable effects on the recombination landscape. There was no association between the inferred recombination rate and base composition, supporting a negligible influence of GC-biased gene conversion in butterflies. We found significant but variable associations between the recombination rate and the density of different classes of transposable elements (TEs), most notably a significant enrichment of SINEs in genomic regions with higher recombination rate. Finally, the analyses unveiled significant enrichment of genes involved in farnesyltransferase activity in recombination cold-spots, potentially indicating that expression of transferases can inhibit formation of chiasmata during meiotic division. Our results provide novel information about recombination rate variation in holocentric organisms and has particular implications for forthcoming research in population genetics, molecular/genome evolution and speciation.

Keywords: Lepidoptera, recombination rate, wood white, linkage disequilibrium, transposable elements, *Leptidea*

Introduction

The meiotic division process allows sexually reproducing organisms to generate haploid gametes. During meiosis, double-strand breaks are induced into DNA, and a proportion (e.g. 5% in *Arabidopsis thaliana*, 10% in *Mus musculus*) of those are resolved as crossovers (Choi and Henderson 2015, Moens et al. 2002), leading to novel combinations of maternal and paternal chromosome segments. It is well established that the frequency and genomic distribution of crossovers can influence both micro- and macro-evolutionary processes – detailed knowledge about the recombination landscape is therefore key for understanding the relative importance of different proximate and ultimate mechanisms affecting genome evolution, generation and maintenance of genetic diversity, adaptation and speciation (Dapper and Payseur 2017, Stapley et al. 2017, Peñalba and Wolf 2020).

The frequency of recombination events resolved as crossovers (from here on recombination) can vary both at the inter- (Stapley et al. 2017, Smukowski and Noor 2011) and intra-specific level (Samuk et al. 2020), as well as between individuals within populations (Johnston et al. 2016) and within individuals over time (Stapley et al. 2017, Peñalba and Wolf 2020). It is well

established that the recombination rate has a genetic component, but the rate can also be influenced by environmental factors (i.e., recombination is partly plastic; Peñalba and Wolf 2020). Recombination rate can also vary considerably across chromosomes and chromosome regions in many species, and mapping this variation may shed light on the mechanistic control of where in the genome recombination is initiated. The reasons for such regional recombination rate variation have been studied in detail in a few organism groups and a handful of consensuses have been reached. First, the size of a chromosome can affect the recombination rate, mainly because correct segregation seems to be dependent on at least one recombination event per chromosome arm in many species (Pardo-Manuel de Villena and Sapienza 2001, Smith and Nambiar 2020). A lower recombination rate on the sex-chromosomes than on the autosomes is also a commonly observed pattern and this difference can potentially be attributed to low sequence homology as a consequence of general degeneration of sex-limited chromosomes (Y or W) (e.g. Bergero and Charlesworth 2009). Within chromosomes, the location of recombination events might be determined by preferential initiation of double-strand breaks close to the telom-

eres and where the chromatin in general is open and accessible (e.g. Haenel et al. 2018, Gray and Cohen 2016). Conversely, the recombination rate is usually suppressed within and around the highly heterochromatic centromeres (Dapper and Payseur 2017, Stapley et al. 2017). Furthermore, physical interference between multiple chiasmata may lead to regional differences in the recombination rate (Gray and Cohen 2016, Peñalba and Wolf 2020). In mammals, the gene PRDM9 mediates recombination by binding to specific sequence motifs, which explains that most recombination events occur in a limited portion of the genome, i.e. recombination hot-spots (e.g. Myers et al. 2005, Grey et al. 2011). However, most vertebrate and all invertebrate lineages lack a functional copy of PRDM9 and recombination initiation must hence be mediated by other factors in these species.

We expect the regional variation in recombination rate to be associated with genomic features, the efficiency of selection and the levels of genetic diversity (Dolgin and Charlesworth 2003, Petrov et al. 2011). Such regional variation in the efficiency of selection can for example affect the distribution of transposable elements (TEs), which tend to accumulate in regions of low recombination rate in both animals (Bartolomé et al. 2002) and plants (Xu and Du 2014). In addition, the frequency of recombination events can also affect nucleotide composition as a direct consequence of GC-biased gene conversion (gBGC), a process that facilitates the fixation of strong (G and C) over weak nucleotides (A and T) during the double-strand break repair step (Duret and Galtier 2009). Although genome-wide estimates of the recombination rate and large-scale variation landscapes have been obtained for many species, detailed recombination maps are still mainly limited to model organisms and domesticated species, and little is known about recombination rate variation in natural populations. Hence, a broader taxonomic sampling will be needed to get a more complete picture of what drives recombination rate variation within genomes and between lineages. This applies not the least to holocentric organisms which have chromosomes without localized centromeres (Suomalainen 1953), such as Lepidoptera (butterflies and moths), where the research on causes and consequences of recombination rate variation is still in its infancy.

Here, we used a large whole-genome re-sequencing data set from a Swedish population of the wood white butterfly (*Leptidea sinapis*) to characterize the fine-scale variation in recombination rate and assess potential associations between recombination rate, nucleotide composition and genomic features. So far, the knowledge about regional recombination rate variation in Lepidoptera is restricted to a handful of pedigree based genetic maps (Davey et al. 2017, Shipilina et al. 2022, Smolander et al. 2022) and this is therefore a spearheading attempt to describe the fine-scale variation in recombination rate and the potential associations with genomic features in a butterfly. The wood white is widely distributed across western Eurasia and shows extreme intraspecific variation in chromosome numbers, with an increasing number of chromosomes in a cline-like pattern from $2n \sim 56 - 60$ in the northern (Scandinavia) and eastern (central Asia) parts of the distribution range to $2n \sim 106 - 110$ in the south-western (Iberia) part of the distribution range (Dincă et al. 2011, Lukhtanov et al. 2018). Hence, some wood white populations differ significantly from the ancestral Lepidoptera chromosome number of $n = 31$ (Robinson 1971, Ahola et al. 2014). Previous studies suggest that recurrent chromosome fissions and fusions underlie this variation (Dincă et al. 2011, Talla et al. 2017) and that at least a handful of fission/fusion polymorphisms seg-

regate in the Swedish population (Höök et al. 2022). The rapid karyotype evolution in wood whites provides a unique system for characterizing the regional variation in recombination rate in a natural population of a holocentric species, and combining it with investigating the effects of segregating chromosome rearrangements on the recombination landscape. Given the key role of recombination in both micro- and macro-evolutionary processes, our results will be important for understanding molecular mechanisms and evolutionary forces affecting genome evolution and divergence processes in holocentric organisms in general.

Results

Demographic history inferences

The demographic trajectories inferred separately for each chromosome jointly revealed the demographic history of the Swedish population of *Leptidea sinapis*. These demographic trajectories shared three main features. Firstly, a maximum effective population size (N_e) of around 106 approximately 10,000 generations before present (BP), preceded by a period of high N_e and a slight increase matching with the end of the Last Glacial Period. After this time point, N_e started to decline exponentially until stabilizing about 100 generations BP. In the most recent past, N_e remained constant. Contemporary estimates of N_e oscillate between 10^3 and 2×10^4 for the different chromosomes (Supplementary Figure 1).

Recombination rate variation and distribution

The estimated genome-wide recombination rate was 7.37 cM/Mb, with measurements for individual chromosomes ranging between 3.5 - 15.3 cM/Mb. There was a marginally non-significant (Spearman $\rho = -0.292$, $p = 0.06$) negative association between the recombination rate and chromosome length (Supplementary Figure 2). Autosomes (7.65 cM/Mb) showed on average a higher recombination rate than the Z-chromosomes (7.03 cM/Mb), although this difference was not significant (Wilcoxon's test, $W = 41$, $p = 0.92$).

Visual inspection of the variation in recombination rate revealed a considerably reduced recombination rate towards chromosome ends (Figure 1A). A formal analysis showed that the recombination rate in the subtelomeres (last 5 100 kb windows at each end of the chromosomes) was significantly reduced (2.46 cM / Mb) compared to the 100 kb windows located in proximal positions of the chromosomes i.e., outside of the subtelomeric regions (Wilcoxon's test, $W = 1,310,856$, $p\text{-value} = 3.18 \times 10^{-66}$). It should be noted that this does not necessarily reflect a low recombination rate in the telomeres, but rather in the subtelomeric regions, since telomeric repeats were not covered in the wood white genome assembly (Höök et al. 2022). We also analyzed each chromosome separately and found a reduction in recombination rate in subtelomeric regions in 50 out of the 58 chromosome ends (29 chromosome pairs, $2n = 58$; Supplementary Table 1). Out of the 8 chromosome ends that did not show a reduced recombination rate, four match observations of segregating fission/fusion polymorphisms – these involve chromosome pairs 18 and 25, 11 and 26, and 5 and 27, respectively (Höök et al. 2022).

Besides a significantly reduced recombination rate in subtelomeric regions in almost all chromosomes, we also found that regional variation differed in other respects. First, some chromosomes showed an obvious unimodal distribution for the

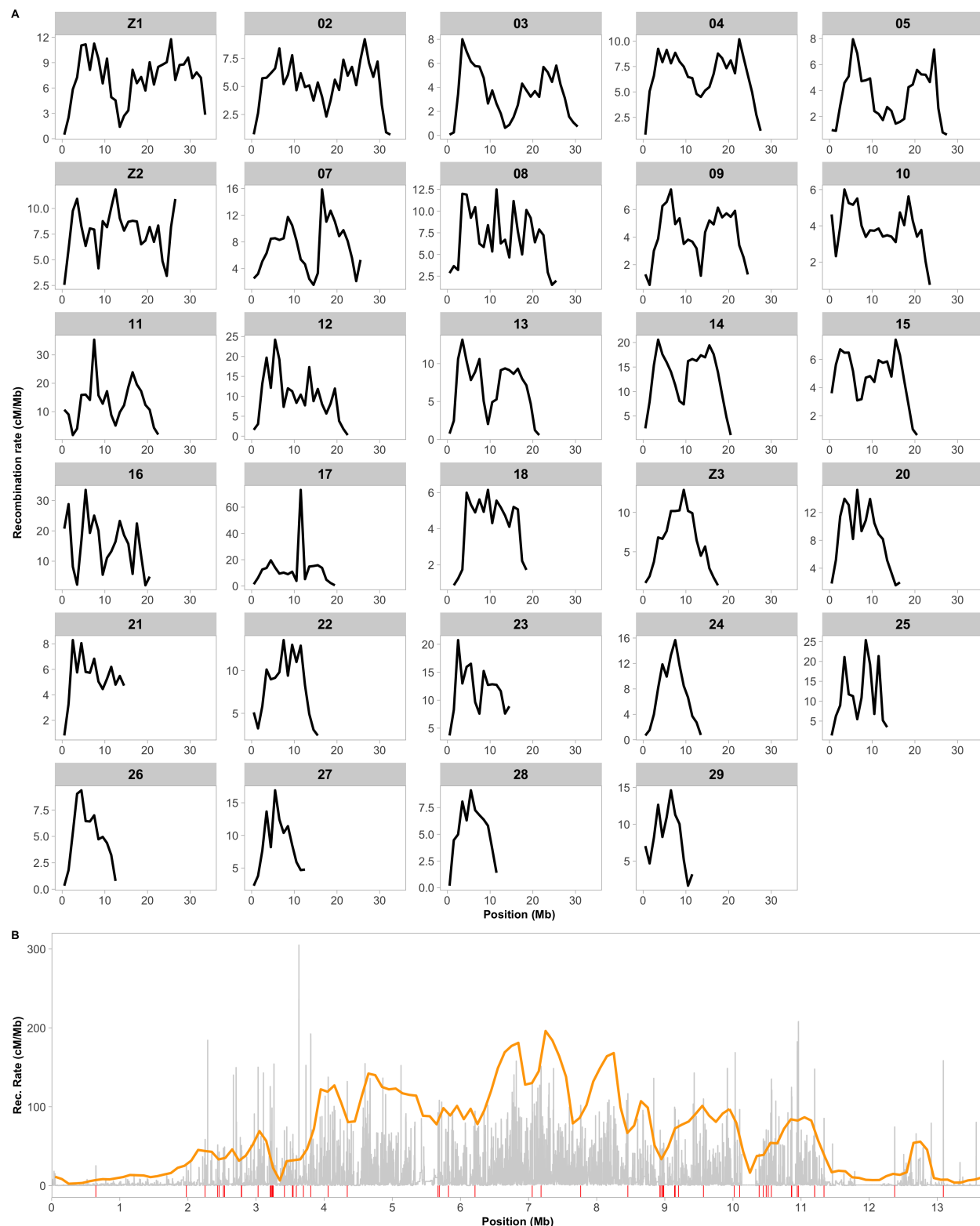


Figure 1 Recombination rate estimations and hot-spot determination. (A) The 1Mb-scale estimates of recombination rate in 1Mb windows are shown in cM/Mb for all chromosomes, ordered by decreasing length. (B) Detailed recombination map for chromosome 24, with all oscillations in recombination rate inferred by pyrro in grey. The orange line represents the 10x background recombination rate, the threshold used to determine the minimum recombination rate of local hot-spots. Red lines underneath the plot indicate the presence of a recombination hot-spot according to our defining parameters. See Materials & Methods for a more detailed description of the parameters.

recombination rate, with a maximum value in the central region and a progressive decrease towards the terminal regions. This was a frequent (but not exclusive) pattern for the shorter chromosomes (Figure 1A). In contrast, the recombination rate was bimodally distributed along some chromosomes, with a central region of reduced recombination in addition to the reduction in subtelomeric regions. This was the most commonly observed pattern for the larger chromosomes (Figure 1).

Recombination hot-spot and cold-spot identification

Informed by coalescent simulations, we developed thresholds for identification recombination hot- and cold spots, which take into account both specific demographic history of the population of interest and the potential stochasticity of our method of recombination rate inference (see methods for further details). Based on these thresholds, a total of 3,124 recombination hot-spots were classified (Figure 1B). The hot-spots had an average length of 1,656 bp and the mean recombination rate within hot-spots was 94.1 ± 62.5 cM/Mb. The highest estimated rate in any hotspot was 708 cM/Mb. The average recombination rate for hot-spots represented an approximate 13-fold increase over the genome-wide recombination rate, but hot-spots only constituted 5.2 Mb (0.87%) of the genome. Recombination hot-spots were found at a significantly lower frequency in the terminal 10% regions of the chromosomes (5% on each end); Wilcoxon's $W = 220.5$; $p = 4.94 \times 10^{-4}$ (Supplementary Figure 3A). A higher density of recombination hot-spots was detected on the three Z-chromosomes (six hot-spots / Mb) when compared to the autosomes (five hot-spots / Mb), but the difference was not significant (Student's t-test, $t = -0.85556$, $df = 27$, $p = 0.20$). Permutation analysis revealed a significantly lower LINE ($p = 0.02$) and LTR density ($p > 0.001$) and a higher SINE density ($p < 0.001$) in the hot-spots, while DNA transposon density did not deviate significantly ($p = 0.49$) from the genome wide average (Figure 2). There was no enrichment of functional gene categories or specific sequence motifs in the hot-spots ($p < 0.05$, Supplementary Table 2).

We also identified 1,283 recombination cold-spots, i.e., regions with considerably reduced recombination rate as compared to the genomic average (see methods). The average length of the cold-spots was 30 kb, and 70 of the cold-spots were longer than 100 kb. Despite the lower frequency of cold-spots than hot-spots, they represented a substantially larger proportion of the genome (38.2 Mb, 6.37%). As expected given the overall reduced recombination rate in subtelomeric regions, cold-spots were particularly abundant in these regions. In particular, subtelomeric regions contained 29.6% of the cold-spots in total, while representing only 4.8% of the genome. The cold-spots located outside subtelomeric regions had an average length of 25.6 kb and 46 were longer than 100 kb. This translates to a significant enrichment in the frequency of cold-spots in the 10% most terminal chromosomal regions compared to the more central positions of the chromosome (Wilcoxon's $W = 3$; $p = 5.62 \times 10^{-4}$) (Supplementary Figure 3B). Similar to the observation for hot-spots, there was no significant difference in cold-spot frequency between the autosomes (1.8 cold-spots / Mb) and the three Z-chromosomes (2.5 cold-spots / Mb) (Student's t-test, $t = -0.74483$, $df = 27$, p -value = 0.231). Permutation analyses revealed a significantly lower transposon density in cold-spots, consistent across all classes ($p < 0.01$; Figure 3). There was also a significant enrichment of genes related to transtransferase activity in the cold-spot regions compared to the genome in general (Supple-

mentary Table 3), but no enrichment of specific sequence motifs in the cold-spots ($p > 0.05$).

Association between recombination rate and base composition

To investigate potential effects of gBGC on the base composition in the wood white genome, we assessed the relationship between the recombination rate and nucleotide composition using a window-based approach. The analysis showed that the recombination rate (averaged over 1 Mb windows) was not significantly correlated with the GC content in the genome in general (Spearman $\rho = -0.07$, $p = 0.10$) (Figure 4A).

Since we observed a significant reduction in recombination rate in subtelomeric regions, we investigated if those regions had deviating base composition. The analysis showed that there was a significantly higher GC content in subtelomeric regions (33.46%) compared to proximal chromosome regions (32.61%) (Wilcoxon's $W = 1,109,810$; $p = 5.73 \times 10^{-22}$).

The GC content was also slightly lower within hot-spots (32.43%) and their 5 kb flanking regions (32.45%) as compared to the genome-wide estimates (32.65%).

Associations between recombination rate and genomic features

We used multiple regression to investigate the relative effect of different explanatory variables (GC content, gene density, DNA transposon, SINE, LINE and LTR retrotransposon densities). The regression model revealed an overall significant association between recombination rate and the explanatory variables ($F(6) = 10.77$, $df = 603$, $p = 2.13 \times 10^{-11}$), but it explained a marginal part of the total variation in the recombination rate ($R^2 = 0.10$, $AdjR^2 = 0.09$) and only SINE and LINE density were significant explanatory variables (Table 1).

The recombination rate was not significantly associated with genome-wide gene density (Spearman $\rho = -0.04$, $p = 0.17$; Figure 4B). However, when partitioning the data and running analyses for different gene elements separately, we observed a significantly lower (Wilcoxon's $W = 1.5971e^{-10}$, p -value $< 2.2 \times 10^{-16}$) recombination rate within exons (5.5 cM/Mb) and in 5' UTR regions (6.3 cM/Mb; Wilcoxon's $W = 1.5869 \times 10^8$, p -value $< 2.2 \times 10^{-16}$) and a significantly higher recombination rate in introns (Wilcoxon's $W = 7.3462 \times 10^{12}$, p -value $< 2.2 \times 10^{-16}$) compared to intergenic regions (7.5 cM/Mb) (Figure 5).

The associations between recombination rate and the densities of different TE classes varied considerably. For all four classes analyzed, we found a significant correlation with the recombination rate, but the direction and strength of these associations varied. DNA transposons (Spearman $\rho = 0.09$, $p = 0.03$) and SINEs (Spearman $\rho = 0.29$, $p = 3.31 \times 10^{-13}$) were positively associated, and LTRs (Spearman $\rho = -0.11$, $p = 8.50 \times 10^{-3}$) and LINEs (Spearman $\rho = -0.19$, $p = 3.01 \times 10^{-6}$) negatively associated with the recombination rate (Figure 6).

Differences between classes of transposable elements were not restricted to the overall associations with recombination rate. The average recombination rate within each class of TEs varied as well. In LINEs (6.6 cM / Mb; Wilcoxon's $W = 5.233 \times 10^{12}$, p -value $< 2.2 \times 10^{-16}$) and DNA transposons (6.8 cM / Mb; Wilcoxon's $W = 2.6104 \times 10^{12}$, p -value $< 2.2 \times 10^{-16}$) the recombination rate was significantly lower, and in SINEs (7.4 cM / Mb; Wilcoxon's $W = 1.2212 \times 10^{12}$, p -value $< 2.2 \times 10^{-16}$) and LTRs (7.8 cM / Mb; Wilcoxon's $W = 7.0468 \times 10^{11}$, p -value $< 2.2 \times 10^{-16}$) the

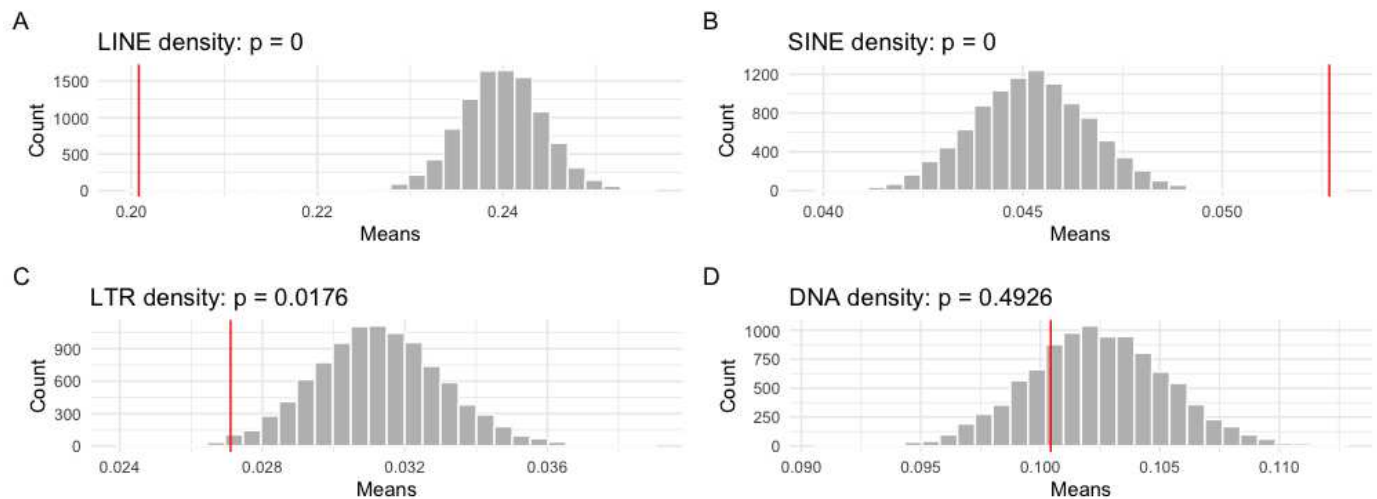


Figure 2 Density of TEs in hot-spots. The bars show the distribution of sampled TE density means for 3,124 random 1,656 bp genomic windows (number and average length of the hot-spots). The red line indicates the observed mean TE densities in the defined hot-spot windows.

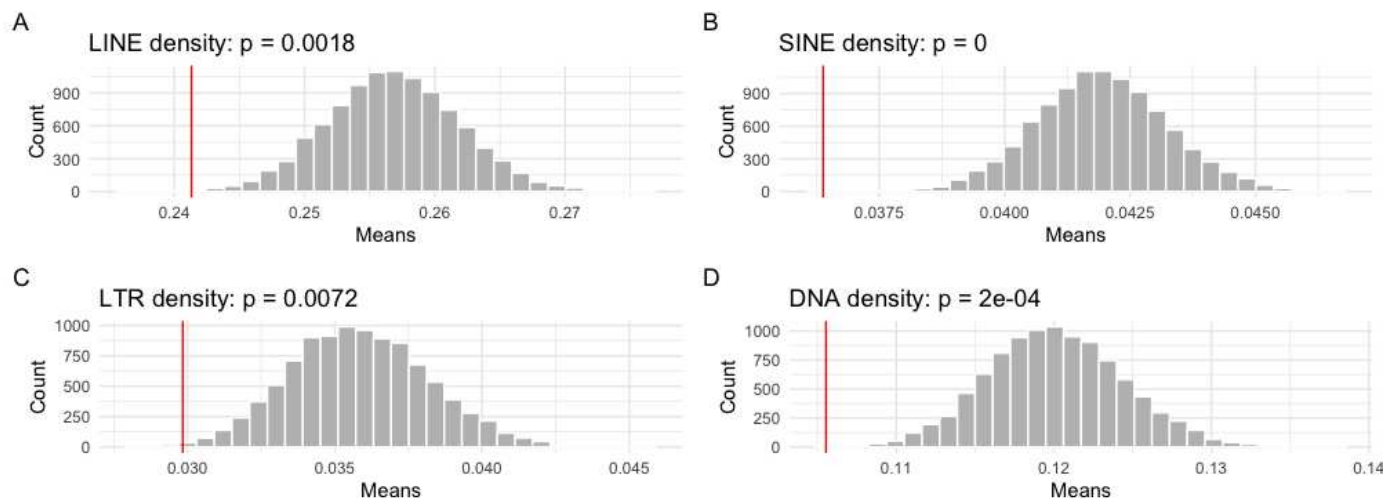


Figure 3 Density of TEs in cold-spots. Histogram shows the distribution of permuted TE density means for 1,283 random 30 kb genomic windows (number and average length of the cold-spots). The red lines indicate the the observed mean TE densities in the defined cold-spot windows.

recombination rate was significantly higher than the genomic average rate (Figure 5).

Discussion

Genome-wide distribution of the recombination rate in wood whites

The genome-wide rate of recombination has been shown to vary considerably between different insect species, from comparatively low in Diptera (< 1 cM / Mb, Beye et al. 2006) to exceptionally high in honeybees (19 cM / Mb; Beye et al. 2006). We found that the genome-wide recombination rate in the Swedish wood white population was 7.37 cM / Mb. This is slightly higher than estimates from other lepidopterans like *Bombyx mori* (4.6 cM / Mb), *Heliconius melpomene* (5.5 cM / Mb) and *H. erato* (6 cM / Mb) (Tobler et al. 2005; Yasukochi 1998; Jiggins et al. 2005) and substantially higher than in vertebrates for which estimates

range between 0.16 cM / Mb in the Atlantic trout to 3.17 cM / Mb in chicken (Beye et al. 2006).

Although not significant at the 5% level, we found that the recombination rate was negatively associated with chromosome size. A clearly deviating rate was found for chromosome 16 (15.3 cM / Mb), which explains why the association between chromosome size and recombination was non-significant. Negative associations between recombination rate and chromosome length have repeatedly been observed in different taxonomic groups, for example yeast (Kaback et al. 1992), humans and rodents (Jensen-Seaman et al. 2004), birds (Backström et al. 2010), cattle (Mouresan et al. 2019) and butterflies (Martin et al. 2019, Shipilina et al. 2022). Such a relationship is expected given that crossovers are necessary for correct segregation of chromosomes during meiosis (Pardo-Manuel de Villena and Sapienza 2001, Smith and Nambiar 2020). Our analysis also showed that longer chromosomes tended to have a bimodal recombination

6 The fine-scale recombination rate variation and associations with genomic features in a butterfly

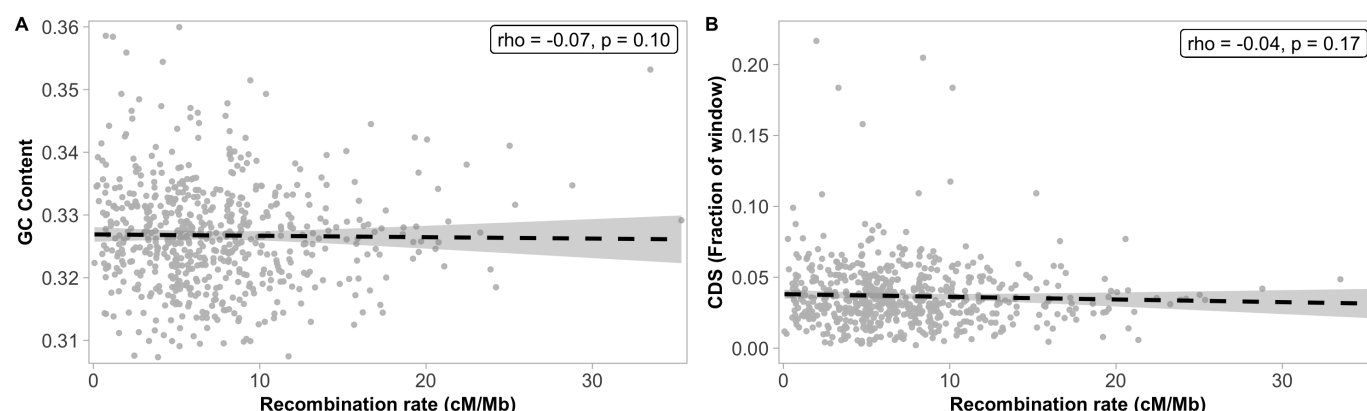


Figure 4 Associations between the recombination rate and base composition (A) and gene density (B). Correlation between parameters was calculated using 1 Mb genomic windows.

Table 1 Linear regression model with six explanatory variables included. Explanatory variables that are significantly associated with recombination rate variation at a 1 Mb scale are highlighted in bold.

	Estimate	Std. error	t	p-value
GC	27.5203	24.6317	1.117	0.264
Gene density	-6.6062	6.7692	-0.976	0.329
DNA	7.2352	16.7373	0.432	0.666
LINE	-30.4439	7.2939	-4.174	$3.44e^{-05}$
SINE	104.8162	16.0648	6.525	$1.45e^{-10}$
LTR	32.1472	23.9983	1.340	0.181

rate distribution with a reduced rate at the chromosome center and towards the chromosome ends. This pattern is in line with findings in other taxa, both organisms with defined centromeres where recombination is reduced (Dapper and Payseur 2017), and holocentric species such as *Caenorhabditis elegans*, in which the recombination rate has been shown to increase with the relative distance from the center of the chromosomes (Prachumwat et al. 2004). The reduced recombination rate in the center of chromosomes in monocentric species is often a direct consequence of the lack of crossing-over events in the centromeres. This explanation is obviously not valid in holocentric lineages. Since the pattern is restricted to larger chromosomes, a potential explanation to the reduced rate in chromosome centers could be the occurrence of multiple chiasmata on larger chromosomes. For example, observations in *Psylla foersteri* suggest that longer chromosomes can accommodate the formation of two simultaneous chiasmata, while shorter chromosomes only have one (Nokkala et al. 2004). In the cases where two chiasmata are formed in a single chromosome, crossover interference may prevent those from forming near each other and tend to drive them towards opposite ends of the chromosome (Otto and Payseur 2019). An additional, but not mutually exclusive explanation, is the formation of the “meiotic bouquet”, a stage in early meiosis characterized by the aggregation of chromosome ends close to the nuclear membrane, which can drive the crossovers towards distal positions and reduce the recombination rate in the center of the larger chromosomes

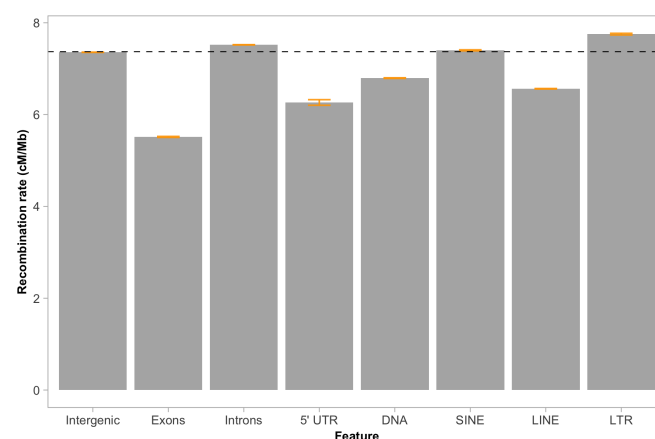


Figure 5 Recombination rate estimates in different gene elements and classes of transposable elements. The genome-wide recombination rate (7.37 cM/Mb) is indicated with the horizontal black dashed line. Orange bars indicate the 95% confidence intervals.

(Scherthan et al. 1996, Haenel et al. 2018).

Besides a reduced recombination rate in the center of larger chromosomes, we also found a reduced recombination rate towards the very ends of chromosomes. This pattern was observed for almost all chromosomes (see exceptions below), irrespective of chromosome size and type. Note that the telomeres, which in Lepidoptera are 6-8 kb long tandem repeats of the motif (TTAGG)_n (Okazaki et al. 1993, Sasaki and Fujiwara 2000) were not assembled in the reference genome. Such decrease in the recombination rate in the subtelomeric regions of the chromosomes has previously been observed in *Heliconius* butterflies (Martin et al. 2019), and also in some other organism groups such as yeast (DuBois et al. 2002, Barton et al. 2008) and flycatchers (Kawakami et al. 2013). A potential explanation for this pattern is that crossover initiation is prevented near the telomeres to minimize the risk for ectopic recombination between non-homologous repeat sequences during meiosis (Smith and Nambiar 2020). As mentioned above, our results showed that the reduced recombination rate towards chromosome ends was not ubiquitous across all chromosomes; eight chromosome ends (one for each of the following chromosomes: 5, 6, 10, 11, 16, 25,

27 and 29) did not show a significant reduction in the recombination rate as compared to each respective intra-chromosomal level. Four of these exceptions (one chromosome end for each of chromosomes 5, 11, 25 and 27) coincide with recently identified fission and fusion polymorphisms segregating in the wood white population in Scandinavia (Höök et al. 2022). These results show that fission/fusion events can have immediate effects on the distribution of crossover events within and between chromosomes.

Characterization of recombination hot-spots and cold-spots

The total number of hot-spots ($n = 3,124$) identified in the wood whites is equivalent to what has previously been observed in for example *Ficedula flycatchers* using a comparable approach (Kawakami et al. 2013). The density of hot-spots was, however, much lower than in humans ($n = 25,000 - 50,000$ hot-spots in an approximately five times larger genome) (Myers et al. 2005). Although the specific thresholds for defining hot-spots vary between studies, they all rely on the comparison of the background and local recombination rates, making the results reasonably comparable. We found that the distribution of hot-spots was similar between wood whites and humans, as hot-spots occurred mostly outside of genes (McVean et al. 2004). This is in contrast to birds, which show an enrichment of hot-spots within genic and regulatory regions (Kawakami et al. 2013; Singhal et al. 2015; Smeds et al. 2016). We found that the frequency of recombination hot-spots and cold-spots was relatively similar in the center of chromosomes, but that the number of hot-spots decreased, and cold-spots were more frequent towards chromosome ends. A higher occurrence of recombination cold-spots in terminal regions of the chromosomes has previously been observed in yeast, which seem to lack crossovers close to chromosome ends altogether (Su et al. 2000; Barton et al. 2003). This observation is also in line with the observed decrease in average recombination rate close to chromosome ends in the wood whites and again suggests that the recombination machinery is partly blocked from accessing the very ends of chromosomes.

We found that the recombination landscape in the wood white was highly variable. This is in line with observations in other organisms like humans, birds (Singhal et al. 2015) and dogs (Axelsson et al. 2012). However, other insects like for example *D. melanogaster* (Comeron et al. 2012), for which detailed recombination maps are available, generally show less pronounced recombination hot-spots. While the hot-spot locations in humans largely are determined by the presence of sequence motifs associated with PRDM9 binding (Grey et al. 2011), little is known about what drives crossovers to occur at specific locations in organisms that lack a functional copy of PRDM9. In order to get preliminary information about potential mechanistic underpinnings of recombination rate variation in wood whites, we therefore assessed if specific sequence motifs or gene categories were enriched in recombination hot-spots and cold-spots. The analyses did not reveal any associations for sequence motifs, and we found no enrichment of specific gene categories in hot-spots. In cold-spots, however, there was a significant enrichment of genes with functions associated to transferase activity. Particularly interesting is the case of farnesyltransferase activity, as farnesylation is a key step for the correct attachment between the spindle and the kinetochores in humans (Moudgil et al. 2015). It is therefore tempting to speculate that the active expression of farnesyltransferase

might block the recombination machinery close to those genes. However, since farnesyltransferases are located in only 13 different cold-spot regions, other forces must also underlie the absence of recombination in many cold-spots.

Associations between recombination, nucleotide composition and gene content

The high-density recombination maps developed here, allowed us to investigate potential associations between the local recombination rate and different genomic features. Such information can be used to deduce the effects of recombination on base composition and/or potential regulatory mechanisms modulating the recombination landscape. A potential driver of a positive association between recombination and nucleotide composition is GC-biased gene conversion (gBGC), i.e., the fixation bias favoring “strong” alleles (G and C) over “weak” alleles (A and T) during meiotic recombination (Duret and Galtier 2009). This process mimics directional selection and can lead to deviating nucleotide composition between regions experiencing different recombination rates. The analysis showed that the local recombination rate was not associated with nucleotide composition (GC-content) in the wood whites. We also found that recombination hot-spots had marginally lower GC-content (the opposite is usually observed when biased fixation is a considerable force; Kawakami et al. 2013). This is in line with a limited effect of gBGC in *Leptidea* butterflies (Boman et al. 2021) and stays in contrast to findings in several other systems like humans (Fullerton et al. 2001; Meunier and Duret 2004), mice (Clément and Arndt 2013), flycatchers (Kawakami et al. 2013) and fruit flies (Marais et al. 2001), as well as plants (Muyle et al. 2011) and yeast (Gerton et al. 2000; Kiktev et al. 2018). As far as we are aware, there are no other studies that have analyzed the strength of gBGC in butterflies outside the *Leptidea* genus. Hence, it is premature to draw conclusions regarding the impact of gBGC on nucleotide composition in Lepidoptera in general.

In many organisms, for example mouse (Paigen et al. 2008) and different plant species (Gaut et al. 2007; Tiley Burleigh 2015), recombination occurs more frequently in gene-dense genomic regions, but we did not find such an association in the wood whites. However, our data showed that the recombination rate was significantly reduced in exons and 5' UTR regions compared to the introns and intergenic regions. This is in line with findings in other insects (Wallberg et al. 2015; Jones et al. 2019), as well as in humans (McVean et al. 2004), where recombination hot-spots mainly occur in the vicinity of, but not within, coding and regulatory regions. The small but significantly elevated recombination rate in introns compared to intergenic regions is consistent with findings in the holocentric nematode *C. elegans* (Prachumwat et al. 2004), but in contrast to the observations in for example *Drosophila* (Carvalho Clark 1999) and humans (Comeron Kreitman 2000). Taken together, this indicates that recombination occurs within genes in butterflies, but that crossovers are partly inhibited in coding sequences which might lead to a slightly elevated rate in introns.

Different TE classes show contrasting association patterns with the recombination rate

Potential associations between recombination and TE densities have mainly been investigated in organisms with defined centromeres (Kent et al. 2017), while investigations in holocentric species are scarce (but see Lavoie et al. 2013, Baril Hayward 2022, Smolander et al. 2022). To investigate the potential asso-

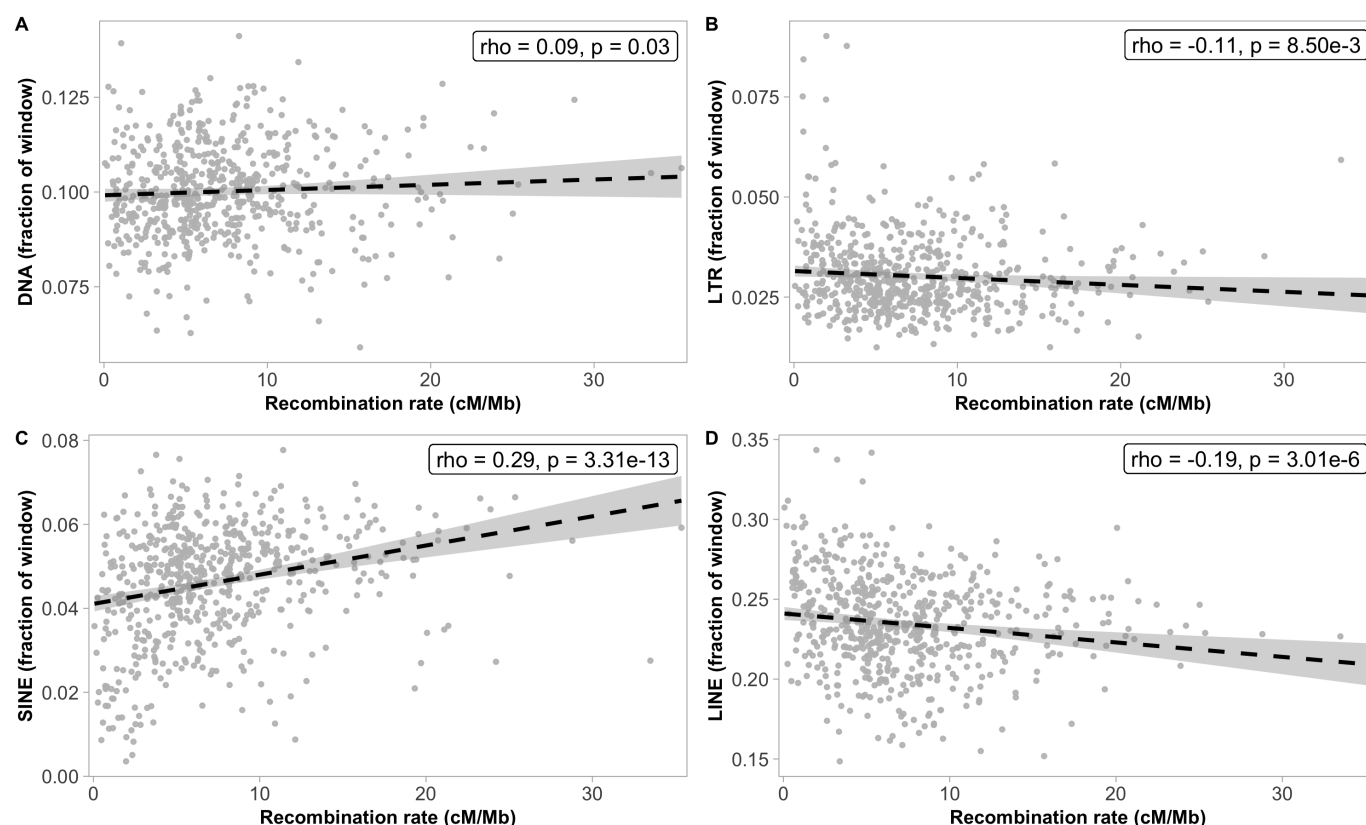


Figure 6 Association between the recombination rate (1 Mb scale) and the proportion of four different TE classes; (A) DNA transposons, (B) LTRs, (C) SINEs and (D) LINEs. The abundance of TEs was calculated as the fraction of each 1 Mb window occupied by each specific element.

ciations between the recombination rate and genomic features in the wood white, we used density information for TEs previously identified in the species (Höök et al. 2022). The analysis revealed that associations between the recombination rate and the abundance of TEs varied considerably depending on the TE class. DNA transposons and SINEs were positively associated, while LINEs and LTRs were negatively associated with the recombination rate. Under the assumption that TE insertions in general are slightly deleterious we would expect a negative correlation between the recombination rate and the abundance of TEs (Kent et al. 2017), as a consequence of more efficient purging of deleterious insertions in regions with a higher recombination rate (Bartolomé et al. 2002, Wright et al. 2003). However, given that recombination is initiated by a double-strand break, it is possible that certain types of TEs are used as a template for the repairing process, driving them to higher frequencies in regions of high recombination rate (Onozawa et al. 2014). SINEs have for example have been shown to use DNA breaks to integrate back into the genome after replication (Singer 1982). Potential associations between TEs and the recombination rate are hence expected to depend on the occurrence of specific classes of TE in the focal study system. For example, Alu elements (a sub-family of SINEs) in humans have been shown to accumulate in regions with elevated recombination (Witherspoon et al. 2009) and SINEs are strongly positively associated with the recombination rate in the painted lady (*Vanessa cardui*) (Shipilina et al. 2022). Similarly, DNA transposons are associated with high recombination rate in *C. elegans* (Duret et al. 2000). The causal-

ity of such associations between the variation in recombination rate and the abundance of TEs is not easy to establish. In cases where TE proliferation has deleterious fitness effects, we expect a negative association between TE abundance and recombination rate. However, presence of Alu elements in humans has been shown to lead to an increase in the local recombination rate, possibly a consequence of that the Alu elements mimic the action of short recombinogenic motifs (Witherspoon et al. 2009). It is also possible that other underlying factors affect the TE and recombination rate distributions similarly – both the TE proliferation and the recombination initiation machinery for example seem to target open chromatin more easily (Kawakami et al. 2013). In the wood whites, the average recombination rate estimates within TE classes did not deviate considerably from the genome-wide average. However, these comparatively minor differences in the recombination rate can indicate differences in the selective pressure against insertion of specific families of TEs. This does not seem to be related with the length of the TEs, as SINEs – which are considerably shorter than LINEs and LTR elements – showed an average recombination rate between the longer types.

In summary, the different TEs showed different associations with the local recombination rate. These results are consistent with findings in other studies and may point toward similar determinants in holocentric organisms compared to those with defined centromeres. However, the causality needs further study, for example by detailed characterization of cross-over regions in large pedigrees.

Materials and methods

Genome assembly

The wood white genome assembly used as reference was developed for another study (Höök et al. 2022). In brief, one mated adult female wood white was caught in Sweden and kept in the lab for egg laying. From the offspring, one male pupa was sampled and flash frozen in liquid nitrogen. The sample was divided to create a 10X Genomics Chromium Genome-library and a Dovetail HiC-library from the same individual. For 10X sequencing, DNA was extracted using a modified HMW salt extraction method (Aljanabi and Martinez 1997). Tissue for HiC-sequencing was disrupted in liquid nitrogen. Library preparations, sequencing and genome assembly was performed by NGI Stockholm. Sequencing was performed on Illumina NovaSeq6000 with a 2x151 setup. 10X linked reads were assembled with 10X Genomics Supernova v2.1.0 (Weisenfeld et al. 2017). HiC reads were processed with Juicer v1.6 (Durand et al. 2016a) and used for scaffolding the 10X assembly with 3DDNA v.180922 (Dudchenko et al. 2017). Resulting assemblies were reviewed with Juicebox v1.11.08 (Durand et al. 2016b). In addition to minor corrections to the initial assembly, two chromosome sized scaffolds were merged. The assembly was finalized with the script 'run-asm-pipeline-post-review.sh' from the 3DDNA pipeline v.180922 (Dudchenko et al. 2017).

Gene annotation

Gene annotation lift-over was performed by aligning wood white protein queries generated by Talla et al. (2017) to the current version of reference genome, using spaln 2.4.0 (Iwata and Ghoto 2012) with the parameters -Q7 -LS -O7 -S3.

TE annotation

Repetitive element consensus sequences were predicted *de novo* using RepeatModeler 1.0.11 (Bao et al. 2015). Transposable elements characterized as unknown were submitted to CENSOR (Bao et al. 2015) for annotation, where any hits with a score < 200 were removed. All predicted sequences were matched against gene annotations using diamond blast 2.0.4 (Buchfink et al. 2021), to correct for annotation errors (bitscore > 100). Transposable elements were then annotated in the genome with RepeatMasker 4.1.0, using the predicted library of consensus sequences in wood white and previously characterized TEs in *Heliconius melpomene* in the RepeatMasker library 4.0.8 (Bao et al. 2015).

Sampling of individuals

Adult male wood whites were collected across the distribution range in Sweden during June and July 2020 (Supplementary Table 4). Sex was determined *in situ* based on two sexually dimorphic characters; the presence of a black apical spot on the forewing and the white coloration of the ventral part of the antennae in males. Sampled individuals were directly preserved in ethanol and frozen at -20°C.

DNA extraction

DNA was extracted following two different protocols. In both cases, the dissected tissue was digested overnight in Laird's buffer and homogenized with 20 µl of proteinase K (20mg/ml, >600 mAU/ml), followed by incubation with RNase A at 37°C for 30 minutes. DNA was extracted from thoraces using salt

extraction; 300 µl of NaCl (5M) was added, followed by centrifugation for 15 minutes at 13,000 revolutions per minute (rpm). Three washing steps were completed with one volume of 70% ethanol and centrifuging for five minutes at maximum speed. The remaining pellet was air-dried and then resuspended in 30 µl of MilliQ H₂O. For the abdomens, a phenol-chloroform extraction protocol was used. Two cycles of phenol:chloroform:isoamyl alcohol (25:24:1) addition and centrifugation for five minutes at 13,000 rpm were completed, plus a third cleaning cycle using only chloroform. Precipitation of DNA was achieved by adding 2x volumes isopropanol + 0.1x 3M NaAc, incubating at -18°C overnight and centrifuging for 15 minutes at 13,000 rpm. The final pellet was resuspended with 30 µl of MilliQ H₂O. DNA purity was assessed with NanoDrop, and concentration measured with Qubit DNA Broad Range.

Sequencing

To capture the genetic variation in the population in Sweden, 84 individuals from different geographic regions and the highest DNA quality were selected for analysis. Library preparation for all 84 samples using the TruSeq PCR-free kit followed by multiplexing, and sequencing on two NovaSeq 6000 S4 lanes with 2x150 bp reads, were performed at the National Genomics Infrastructure (NGI), Stockholm.

Read trimming

Illumina sequencing adapters were trimmed by eliminating the first fifteen base pairs (bp) on each end of the raw reads with CutAdapt 1.9.1 (Martin 2011), filtered on Q-score < 30 and a minimum length of 30 bp. Read quality after cleaning was assessed with FastQC (Andrews 2010). Before filtering, an average of 4.3 million reads per sample were obtained, and 2.5% were filtered out.

Mapping and filtering

For each individual, paired-end reads were mapped to the reference genome with bwa v0.7.17 (Li and Durbin 2009). Samtools v1.10 (Li et al. 2009) was used to select reads with paired information. MarkDuplicatesSpark as implemented in GATK v4.1.4.1 (McKenna et al. 2010) was used to eliminate duplicated regions with the -remove-sequencing-duplicates option.

Variant calling and filtering

The tool HaplotypeCaller in GATK v4.1.4.1 (McKenna et al. 2010) was used for variant calling. Each chromosome for each individual was processed separately, and the resulting 84 files for each chromosome were grouped and converted into a VCF file with the GATK v4.1.4.1 tools Combine_gVCF and Genotype_gVCF (McKenna et al. 2010), respectively. Total variant count was obtained with the stats option in bcftools v1.10 (Li 2011). The variants were filtered to have a minimum minor allele count (MAC) of two, a per-site depth between 10 and 50, minimum per site quality of 30, and < 20% per-site missing data with vcftools 0.1.15 (Danecek et al. 2011). Additionally, all insertions and deletions were removed with the -remove-indels option. The number of remaining sites in each chromosome was counted again with the stats option in bcftools v1.10 (Li 2011). Initial variant calling resulted in a total of 51,189,479 markers along the genome (11,055,543 indels + 40,133,936 SNPs), of which 10,565,404 SNPs remained after filtering. This represents a genome-wide average of ~17.6 SNPs/kb, given the ~ 600 Mb total length of the reference genome.

Inference of the demographic history

SMC++ (Terhorst et al. 2017) was used to infer the demographic history of each individual chromosome, using a set of six “distinguished” individuals. These six individuals were selected among the 10 with a higher average sequencing coverage for the variants after filtering, so that they constituted a good representation of the geographic distribution of the species. The per-base mutation rate was set at 2.9×10^{-9} per generation, an estimate based on mutation frequency in *H. melpomene* (Keightley et al. 2015). Known invariable regions such as centromeres must be masked before inferring the demographic history, as they can interfere with the signal. Since Lepidoptera are holocentric and lack defined centromeres, a cut-off value of 150 kb was set instead, so that any longer invariable region was considered as missing data and discarded for the demographic inference. The demographic trajectories were inferred for the last 5 million generations, as defaulted by the program.

Recombination rate estimation

The chromosome-specific demographic trajectories were used together with the VCF files to obtain high-resolution recombination maps using pyrro (Spence and Song 2019). An algorithm implemented in the software LDpop (Kamm et al. 2016) was used to compute a table of two-loci likelihoods under the coalescent with recombination using the chromosome-specific demographic trajectories as input. The same mutation rate as before was used, together with the parameter values $n = 168$ (twice the number of diploid individuals), and $N = 210$ (25% larger than n , as recommended by the manual). A relative tolerance (`--decimate_rel_tol`) value of 0.1 was used, together with the `--approx` flag, recommended for large datasets. Different window sizes (maximum distance between SNPs) and the block penalty (determinant of the smoothness of the curve) were tested for each chromosome, and the most appropriate were selected based on the correlation between the data in the likelihood tables and simulated data at different scales (1 bp, 10 kb and 100 kb). For all chromosomes, the best block_penalty was 25, and the best suitable window_size ranged between 50 and 100. The look-up table, together with the final VCF file for each chromosome were used to infer the local recombination rate using the most appropriate parameter values and the `--fast_missing` flag. The per-base, per-generation recombination rate between each pair of SNP markers was obtained in the end. A positive association between the recombination rate and marker density was observed at a 1Mb scale (see Supplementary Information and Supplementary Figure 4 for further discussion).

Distribution of recombination rate variation and identification of hot-spots and cold-spots

Regional recombination rate estimates were obtained in windows of two different sizes (100 kb and 1 Mb) with a custom script by calculating the weighted mean on each interval between markers, accounting for their length and recombination rate. We used simulation-based approach to establish thresholds for hot-spot identification. By performing coalescent simulation with the flat recombination landscape but taking into account previously inferred population history and genetic drift and we obtained levels of recombination rate variation, which are due to our imputation strategy. msprime 1.1.1 (Kelleher et al. 2016) was used to simulate 99 independent sequences, each 100 kb long, and the VCF files for each sequence were concatenated. The parameters for the simulations included a flat recombination

landscape (7.37 cM/Mb, same as the obtained genome-wide rate) and the same mutation rate used for SMC++, together with a demographic trajectory that approximately reflects the inferred trajectories from the empirical data; exponential growth in the interval 104-106 generations BP, exponential decline 102-104 generations BP, and stable population in the last 102 generations. At each time point, N_e was calculated as the average of the estimates across chromosomes. The simulated genomic sequences were analyzed in pyrro according to the steps described earlier. The resulting regional recombination rates predominantly oscillated in the range 2-8 cM/Mb, with a maximum of 20 cM/Mb, a 3-fold higher rate compared to the genome-wide average (Supplementary Figure 5). Recombination peaks occurring at boundaries between independently obtained VCF files were omitted. Therefore, simulation allowed us to establish lower bound for hot-spot threshold. For the final (more conservative) threshold we choose regions with a recombination rate higher than 25 cM/Mb, between 750 and 10,000 bp long, and showing a 10-fold increase over the regional background recombination rate (the mean rate in the focal 100 kb window and the two flanking windows, in total 300 kb). To avoid biases resulting from erroneously called variants, hot-spots that included less than four markers (i.e., three intervals) were discarded. Recombination cold-spots were identified as regions with a recombination rate 10-fold lower than the genome-wide average, and including 4 markers with no length limitations.

For the assessment of the distribution of recombination hot-spots and cold-spots, we defined the subtelomeric regions as the last 500 kb on each chromosome end. The rest of the genome was considered proximal. Flanking regions to recombination hot-spots were defined as the 5 kb segments on each side of each accepted hot-spot.

Association of recombination rate with GC content, gene density and TE classes

A multiple linear regression model was constructed to disentangle the explanatory potential of each genomic feature in the variation of recombination rate at a 1 Mb scale, using the `lm` function in base R (R Core Team 2021). The linear model had the recombination rate as response variable, and the explanatory variables included the GC content, the gene density, and the relative abundance of four TE classes (DNA transposons, and LINE, SINE and LTR retrotransposons). Potential associations between the regional recombination rate and the different genomic features were also analyzed with Spearman's rank correlation tests using the `corr` option in base R (R Core Team 2021).

BEDTools 2.29.2 (Quinlan and Hall 2010) `maskfasta` option was used to select all annotated exons and TEs. Base composition for specific regions (corrected for masked positions) was obtained with BEDTools `nuc` option. Gene and TE densities were calculated as the proportion of a region covered by the annotated sequences of each category.

To assess potential variation in the recombination rate within specific genomic features, we also estimated the recombination rate within each TE class, different regions in protein coding genes (exons, introns and upstream regions) and intergenic sequence. Since 5' UTR regions were not included in the annotation file, we used the 100 bp upstream of the first exon of each gene – a conservative selection to represent the 5' UTR (Chen et al. 2011, 2014). In order to avoid biases due to differential selective pressures, only the intervals between markers from the pyrro output that were positioned completely within the feature

were considered - i.e. for a given exon, if an interval between markers overlapped completely with the predicted exon, it was retained; if the overlap was only partial (including part of the exon sequence but also part of a neighboring intron or UTR), it was discarded. This overlap was checked with BEDTools 2.29.2 (Quinlan and Hall 2010) intersect option using the flag -f 1.

Permutation test to assess TE density in the cold- and hot-spots

To assess the TE density in the cold and hotspot regions we ran a permutation test in R v4.2.1 (R Core Team, 2013), by resampling 10,000 means of windows reflecting the count and average length of the outlier windows and assessing where in this distribution of resampled means the value of the hot- and cold-spot windows appears. We then performed a two-sided test by assessing the number of means that exceeded the original difference.

Gene Set Enrichment Analysis (GSEA) for the genes present in the cold- and hot-spot regions We assessed enrichment of functional categories in the cold and hot spot regions using topGO v2.44.0 (Alexa and Rahnenfuhrer 2021) in R v4.2.1 (R Core Team, 2013). We used the annotated gene set with gene ontology (GO) terms associated to molecular function. To assess significance, we used the Fisher's exact test and the default algorithm ("weight01") accounting for the hierarchical structure of the GO-terms (Alexa et al. 2006). We adjusted the p-values with Benjamini-Hochberg's method of multiple test correction (p.adjust(x, method = "fdr")). We used HOMER v4.11 (Heinz et al. 2010) to assess motif enrichment in the cold and hot spot windows.

Data access

All raw data generated in this study has been submitted to the European Nucleotide Archive (ENA; <https://www.ebi.ac.uk/ena/browser/home>) under accession number PRJEB56690. Command lines are available on the group's GitHub page (<https://github.com/EBC-butterfly-genomics-team>).

Competing Interests Statement

The authors declare that they have no conflict of interest.

Acknowledgements

This work was supported by a research grant from the Swedish Research Council (Vetenskapsrådet Grant ID: 2019-04791) to N.B. The authors acknowledge support from the National Genomics Infrastructure in Stockholm funded by Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council, and SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. This work was also supported by NBIS/SciLifeLab long-term bioinformatics support (WABI). R.V. was supported by grant PID2019-107078GB-I00, funded by MCIN/AEI/10.13039/501100011033.

References

Ahola V, Lehtonen R, Somervuo P, Salmela L, Koskinen P, Rastas P, Välimäki N, Paulin L, Kvist J, Wahlberg N et al. 2014. The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nature Communications* 5: 4737. doi:10.1038/ncomms5737.

Aljanabi SM, Martinez I. 1997. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research* 25: 4692-4693.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.

Axelsson E, Webster MT, Ratnakumar A, The LUPA Consortium, Ponting CP, Lindblad-Toh K. 2012. Death of PRDM9 coincides with stabilization of recombination landscape in the dog genome. *Genome Research* 22: 51 - 63.

Backström N, Forstmeier W, Schielzeth H, Mellenius H, Nam K, Bolund E, Webster MT, Öst T, Schneider M, Kempnaers B. et al. 2010. The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Research* 20: 485-495.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6: 11. doi:10.1186/s13100-015-0041-9.

Baril T, Hayward A. 2022. Migrants within migrants: exploring transposable element dynamics in the monarch butterfly, *Danaus plexippus*. *Mobile DNA* 13: 5. doi:10.1186/s13100-022-00263-5.

Bartolomé C, Maside X, Charlesworth B. 2002. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. *Molecular Biology and Evolution* 19: 926-937.

Barton AB, Su Y, Lamb J, Barber D, Kaback DB. 2003. A function for subtelomeric DNA in *Saccharomyces cerevisiae*. *Genetics* 165: 929 - 934.

Barton AB, Pekosz MR, Kurvathi RS, Kaback DB. 2008. Meiotic recombination at the ends of chromosomes in *Saccharomyces cerevisiae*. *Genetics* 179: 1221 - 1235.

Bergero R, Charlesworth D. 2009. The evolution of restricted recombination in sex chromosomes. *Trends in Ecology and Evolution* 24: 94 - 102.

Beye M, Gattermeier I, Hasselmann M, Gempe T, Schioett M, Baines JF, Schlupalius D, Mougél F, Emore C, Rueppell O et al. 2006. Exceptionally high levels of recombination across the honey bee genome. *Genome Research* 16: 1339 - 1344.

Boman J, Mugal CF, Backström N. 2021. The effects of GC-biased gene conversion on patterns of genetic diversity among and across butterfly genomes. *Genome Biology and Evolution* 13: evab064. doi:10.1093/gbe/evab064.

Buchfink B, Reuter K, Drost HG. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* 18: 366 - 368.

Carvalho A, Clark A. 1999. Intron size and natural selection. *Nature* 401: 344.

Chen C, Lin H, Pan C, Chen F. 2011. The length evolution of 5' untranslated regions - the stochastic model revisited. *IEEE International Conference on Bioinformatics and Engineering*: 162 - 166.

Chen X, Lü X, Zhang Y. 2014. Characterization of protein phosphatase 5 from three Lepidopteran insects: *Helicoverpa armigera*, *Mythimna separata* and *Plutella xylostella*. *PLoS One* 9: e97437. doi:10.1371/journal.pone.0097437

Choi K, Henderson IR. 2015. Meiotic recombination hot-spots - a comparative view. *Plant Journal* 83: 52 - 61.

Clément Y, Arndt PF. 2013. Meiotic recombination strongly influences GC-content evolution in short regions in the mouse genome. *Molecular Biology and Evolution* 30: 2612 - 2618.

Cameron JM, Kretman M. 2000. The correlation between intron length and recombination in *Drosophila*: dynamic equilibrium between mutational and selective forces. *Genetics* 156: 1175 - 1190.

Cameron JM, Kreitman M. 2002. Population, evolutionary and genomic consequences of interference selection. *Genetics* 161: 389 - 410.

Cameron JM, Ratnappan R, Bailin S. 2012. The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genetics* 8: e1002905. doi:10.1371/journal.pgen.1002905

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics* 27: 2156 - 2158.

Dantart J. 2003. *Leptidea sinapis* and *L. reali*, sibling species as yet poorly known in Catalonia. *Cynthia: Bulletin of the Catalan Butterfly Monitoring Scheme* 3: 5 - 6.

- Dapper AL, Payseur BA. 2017. Connecting theory and data to understand recombination rate evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* 372: 1 – 11. doi:10.1098/rstb.2016.0469.
- Davey JW, Barker SL, Rastas PM, Pinharanda A, Martin SH, Durbin R, McMillan WO, Merrill RM, Jiggins CD. 2017. No evidence for maintenance of a sympatric *Heliconius* species barrier by chromosomal inversions. *Evolution Letters* 1: 138 – 154.
- Dincă V, Lukhtanov VA, Talavera G, Vila R. 2011. Unexpected layers of cryptic diversity in wood white *Leptidea* butterflies. *Nature Communications* 2: e324. doi:10.1038/ncomms1329.
- Dolgin ES, Charlesworth B. 2008. The effects of recombination rate on the distribution and abundance of transposable elements. *Genetics* 178: 2169 – 2177.
- DuBois ML, Haimberger ZW, McIntosh MW, Gottschling DE. 2002. A quantitative assay for telomere protection in *Saccharomyces cerevisiae*. *Genetics* 161: 995 – 1013.
- Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 356: 92 – 95.
- Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. 2016a. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* 3: 95 – 98.
- Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. 2016b. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems* 3: 99 – 101.
- Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics* 10: 285 – 311.
- Duret L, Marais G, Biémont C. 2000. Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*. *Genetics* 156: 1661 – 1669.
- Fang L, Ye J, Li N, Zhang Y, Li S, Wong G, Wang J. 2008. Positive correlation between recombination rate and nucleotide diversity is shown under domestication selection in the chicken genome. *China Science Bulletin* 53: 746 – 750.
- Friberg M, Bergman M, Kullberg J, Wahlberg N, Wiklund C. 2008. Niche separation in space and time between two sympatric sister species – a case of ecological pleiotropy. *Evolutionary Ecology* 22: 1 – 18.
- Fullerton SM, Bernardo Carvalho A, Clark AG. 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Molecular Biology and Evolution* 18: 1139 – 1142.
- Gaut B, Wright S, Rizzon C, Dvorak J, Anderson LK. 2007. Recombination: an underappreciated factor in the evolution of plant genomes. *Nature Reviews Genetics* 8: 77 – 84.
- Gerton JL, DeRisi J, Shroff R, Lichten M, Brown PO, Petes TD. 2000. Global mapping of meiotic recombination hot-spots and cold-spots in the yeast *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences* 97: 11383 – 11390. doi:10.1073/pnas.97.21.11383.
- Gray S, Cohen PE. 2016. Control of meiotic crossovers: from double-strand break formation to designation. *Annual Reviews Genetics* 50: 175 – 210.
- Grey C, Barthès P, Chauveau-Le Fric G, Langa F, Baudat F, De Massy B. 2011. Mouse PRDM9 DNA-binding specificity determines sites of histone H3 lysine 4 trimethylation for initiation of meiotic recombination. *PLoS Biology* 9: e1001176. doi:10.1371/journal.pbio.1001176.
- Günther T, Malmström H, Svensson EM, Omrak A, Sánchez-Quinto F, Kılınc GM, Krzewinska M, Eriksson G, Fraser M, Edlund H et al. 2018. Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biology* 16: e2003703. doi:10.1371/journal.pbio.2003703.
- Haanel Q, Laurentino TG, Roesti M, Berner D. 2018. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. *Molecular Ecology* 27: 2477–2497.
- Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* 38:576 – 589.
- Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genetic Research* 8: 269 – 294.
- Höök L, Näsvalk K, Vila R, Wiklund C, Backström N. 2022. High-density linkage maps and chromosome level genome assemblies unveil direction and frequency of extensive structural rearrangements in wood white butterflies (*Leptidea* sp.). [Submitted]
- Iwata H, Gotoh O. 2012. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Research* 40: e161. doi:10.1093/nar/gks708.
- Jacobs GS, Sluckin TJ, Kivisild T. 2016. Refining the use of linkage disequilibrium as a robust signature of selective sweeps. *Genetics* 203: 1807 – 1825.
- Jeffcoate S, Joy J. 2011. Evidence-based national recovery plan for *Leptidea sinapis* (wood white butterfly) in Southern Britain. *Journal of Insect Conservation* 15: 759 – 763.
- Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome research* 14: 528 – 538.
- Jiggins CD, Mavarez J, Beltrán M, McMillan WO, Spencer Johnston J, Bermingham E. 2005. A genetic linkage map of the mimetic butterfly *Heliconius melpomene*. *Genetics* 171: 557 – 570.
- Johnston SE, Berenos C, Slate J, Pemberton JM. 2016. Conserved genetic architecture underlying individual recombination rate variation in a wild population of soay sheep (*Ovis aries*). *Genetics* 203: 583 – 598.
- Jones JC, Wallberg A, Christmas MJ, Kapheim KM, Webster MT. 2019. Extreme differences in recombination rate between the genomes of a solitary and a social bee. *Molecular Biology and Evolution* 36: 2277 – 2291.
- Kaback DB, Guacci V, Barber D, Mahon JW. 1992. Chromosome size-dependent control of meiotic recombination. *Science* 256: 228 – 232.
- Kamm JA, Spence JP, Chan J, Song YS. 2016. Two-locus likelihoods under variable population size and fine-scale recombination rate estimation. *Genetics* 203: 1381 – 1399.
- Kawakami T, Mugal CF, Suh A, Nater A, Burri R, Smeds L, Ellegren H. 2013. Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. *Molecular Ecology* 26: 4158 – 4172.
- Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, Mallet J, Davey JW, Jiggins CD. 2015. Estimation of the spontaneous mutation rate in *Heliconius melpomene*. *Molecular Biology and Evolution* 32: 239 – 243.
- Kelleher J, Etheridge AM, McVean G. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology* 12: e1004842. doi:10.1371/journal.pcbi.1004842.
- Kent TV, Uzunović J, Wright SI. 2017. Coevolution between transposable elements and recombination. *Philosophical Transactions of the Royal Society B* 372: 20160458. doi:10.1098/rstb.2016.0458.
- Kiktev DA, Sheng Z, Lobachev KS, Petes TD. 2018. GC content elevates mutation and recombination rates in the yeast *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences* 115: E7109 – E7118. doi: 10.1073/pnas.1807334115.
- Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167: 1513 – 1524.
- Lavoie CA, Platt RN, Novick PA, Counterman BA, Ray DA. 2013. Transposable element evolution in *Heliconius* suggests genome diversity within Lepidoptera. *Mobile DNA* 4: 21. doi:10.1186/1759-8753-4-21.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* 25: 1754 – 1760.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25: 2078 – 2079.
- Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27: 2987 – 2993.
- Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475: 493 – 496.

- Lin Z, Li W. 2011. Evolution of 5' Untranslated region length and gene expression reprogramming in yeasts. *Molecular Biology and Evolution* 29: 81 – 89.
- Lukhtanov VA, Dincă V, Friberg M, Šíchová J, Olofsson M, Vila R, Marec F, Wiklund C. 2018. Versatility of multivalent orientation, inverted meiosis, and rescued fitness in holocentric chromosomal hybrids. *Proceedings of the National Academy of Sciences of the United States of America* 115: E9610 – E9619. doi:10.1073/pnas.1802610115.
- Marais G, Charlesworth B. Genome evolution: recombination speeds up adaptive evolution. *Current biology* 13: R68 – R70.
- Marais G, Mouchiroud D, Duret L. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. *Proceedings of the National Academy of Sciences* 98: 5688 – 5692. doi:10.1073/pnas.091427698.
- Margus A, Lindström L. 2020. Prolonged diapause has sex-specific fertility and fitness costs. *Evolutionary Ecology* 34: 41 – 57.
- Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 17: 10 – 12.
- Martin SH, Davey JW, Salazar C, Jiggins CD. 2019. Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biology* 17: e2006288. doi:10.1371/journal.pbio.2006288.
- Mather N, Traves SM, Ho SYW. 2020. A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data. *Ecology and Evolution* 10: 579 – 589.
- Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genome Research* 23: 23 – 35.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernyt-sky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20: 1297 – 1303.
- McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* 304: 581 – 584.
- Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Molecular Biology and Evolution* 21: 984 – 990.
- Moens PB, Kolas NK, Tarsounas M, Marcon E, Cohen PE, Spyropoulos B. 2002. The time course and chromosomal localization of recombination-related proteins at meiosis in the mouse are compatible with models that can resolve the early DNA-DNA interactions without reciprocal recombination. *Journal of Cell Science* 115: 1611-1622.
- Moudgil DK, Westcott N, Famulski JK, Patel K, Macdonald D, Hang H, Chan GKT. 2015. A novel role of farnesylation in targeting a mitotic checkpoint protein, human Spindly, to kinetochores. *Journal in Cell Biology* 208: 881 – 896.
- Mouresan EF, González-Rodríguez A, Cañas-Álvarez JJ, Munilla S, Altarriba J, Díaz C, Baró JA, Molina A, López-Buesa P, Piedrafita J et al. 2019. Mapping recombination rate on the autosomal chromosomes based on the persistency of linkage disequilibrium phase among autochthonous beef cattle populations in Spain. *Frontiers in Genetics* 10: 1170. doi:10.3389/fgene-2019.01170.
- Muyle A, Serres-Giardi L, Ressayre A, Escobar J, Glémin S. 2011. GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). *Molecular Biology and Evolution* 28: 2695 – 2706.
- Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hot-spots across the human genome. *Science* 310: 321 – 324.
- Nokkala S, Kuznetsova VG, Maryanska-Nadachowska A, Nokkala C. 2004. Holocentric chromosomes in meiosis. I. Restriction of the number of chiasmata in bivalents. *Chromosome Research* 12: 733 – 739.
- Okazaki S, Tsuchida K, Maekawa H, Ishikawa H, Fujiwara H. 1993. Identification of a pentanucleotide telomeric sequence, (TTAGG)_n, in the silkworm *Bombyx mori* and in other insects. *Molecular and Cellular Biology* 13: 1424 – 1432.
- Otto SP, Payseur BA. 2019. Crossover interference: shedding light on the evolution of recombination. *Annual Review of Genetics* 53: 19 – 44.
- Onozawa M, Zhang Z, Kim YJ, Goldberg L, Varga T, Leif Bergsagel P, Kuehl WM, Aplan PD. 2014. Repair of DNA double-strand breaks by templated nucleotide sequence insertions derived from distant regions of the genome. *Proceedings of the National Academy of Sciences* 111: 7729 – 7734.
- Paigen K, Szatkiewicz JP, Sawyer K, Leahy N, Parvanov ED, Ng SH, Graber JH, Broman KW, Petkov PM. 2008. The recombinational anatomy of a mouse chromosome. *PLoS Genetics* 4: e1000119. doi:10.1371/journal.pgen.1000119.
- Pardo-Manuel de Villena F, Sapienza C. 2001. Recombination is proportional to the number of chromosome arms in mammals. *Mammalian Genome* 12: 318 – 322.
- Peñalba JV, Wolf JBW. 2020. From molecules to populations: appreciating and estimating recombination rate variation. *Nature Reviews Genetics* 21: 476 – 492.
- Petrov DA, Fiston-Lavier AS, Lipatov M, Lenkov K, González J. 2011. Population genomics of transposable elements in *Drosophila melanogaster*. *Molecular Biology and Evolution* 28: 1633 – 1644.
- Powell JA. 1989. Synchronized, mass-emergence of a *Yucca* moth, *Proxodus Y-inversus* (Lepidoptera: Prodoxidae) after 16 and 17 years in diapause. *Oecologia* 81: 490 – 493.
- Prachumwat A, DeVincentis L, Palopoli MF. 2004. Intron size correlates positively with recombination rate in *Caenorhabditis elegans*. *Genetics* 166: 1585 – 1590.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841 – 842.
- Robinson R. 1971. *Lepidoptera genetics*. 1st ed. Pergamon Press, Oxford.
- R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at: <https://www.R-project.org/>.
- Samuk K, Manzano-Winkler B, Ritz KR, Noor MAF. 2020. Natural selection shapes variation in genome-wide recombination rate in *Drosophila pseudoobscura*. *Current Biology* 30: 1517 – 1528.
- Sasaki T, Fujiwara H. 2000. Detection and distribution patterns of telomerase activity in insects. *European Journal of Biochemistry* 267: 3025 – 3031.
- Scherthan H, Weich S, Schwegler H, Heyting C, Härle M, Cremer T. 1996. Centromere and telomere movements during early meiotic prophase of mouse and man are associated with the onset of chromosome pairing. *Journal of Cell Biology* 134: 1109 – 1125.
- Shen C, Li X, Zhang R, Lin Z. 2017. Genome-wide recombination rate variation in a recombination map of cotton. *PLoS ONE* 12: e0188682. doi:10.1371/journal.pone.0188682.
- Shipilina D, Näsvalk K, Höök L, Vila R, Talavera G, Backström N. 2022. Linkage mapping and genome annotation give novel insights into gene family expansions and regional recombination rate variation in the painted lady (*Vanessa cardui*) butterfly. *Genomics* 114: 110481. doi:10.1016/j.ygeno.2022.110481
- Singer MF. 1982. SINES and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. *Cell* 28: 433 – 434.
- Singhal S, Leffler EM, Sannareddy K, Turner I, Venn O, Hooper DM, Strand AI, Li Q, Raney B, Balakrishnan CN et al. 2015. Stable recombination hot-spots in birds. *Science* 350: 928 – 932.
- Smeds L, Mugal CF, Qvarnström A, Ellegren H. 2016. High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. *PLOS Genetics* 12: e1006044. doi:10.1371/journal.pgen.1006044.
- Smith GR, Nambiar M. 2020. New solutions to old problems: mechanisms of meiotic crossover control. *Trends in Genetics* 36: 337 – 346.
- Smolander O-P, Blande D, Ahola V, Rastas P, Tanskanen J, Kammonen JI, Oostra V, Pellegrini L, Ikonen S, Dallas T, et al. 2022. Improved chromosome-level genome assembly of the Glanville fritillary butterfly (*Melitaea cinxia*) integrating Pacific Biosciences long reads and a high-density linkage map, GigaScience 11: giab097. doi:10.1093/gigascience/giab097.
- Smukowski CS, Noor MA. 2011. Recombination rate variation in closely related species. *Heredity* 107: 496 – 508.
- Spence JP, Song YS. 2019. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances* 5: eaaw9206. doi:10.1126/sciadv.aaw9206.
- Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM. 2017. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Philosophical Transactions of the Royal Society B* 372: 20160455. doi:10.1098/rstb.2016.0455.

14 The fine-scale recombination rate variation and associations with genomic features in a butterfly

Su Y, Barton AB, Kaback DB. 2000. Decreased meiotic reciprocal recombination in subtelomeric regions in *Saccharomyces cerevisiae*. *Chromosoma* 109: 467 – 475.

Suomalainen E. 1953. The kinetochore and the bivalent structure in the Lepidoptera. *Hereditas* 39: 88 – 96.

Talla V, Suh A, Kalsoom F, Dincă V, Vila R, Friberg M, Wiklund C, Backström B. 2017. Rapid increase in genome size as a consequence of transposable element hyperactivity in wood-white (Leptidea) butterflies. *Genome Biology and Evolution* 9: 2491 – 2505.

Talla V, Johansson A, Dinca V, Vila R, Friberg M, Wiklund C, Backström N. 2019. Lack of gene flow: Narrow and dispersed differentiation island in a triplet of Leptidea butterfly species. *Molecular Ecology* 28: 3756 – 3770.

Terhorst J, Kamm J, Song Y. 2017. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics* 49: 303 – 309.

Tiley GP, Burleigh JG. 2015. The relationship of recombination rate, genome structure, and patterns of molecular evolution across angiosperms. *BMC Evolutionary Biology* 15: 194. doi:10.1186/s12862-015-0473-3.

Tobler A, Kapan D, Flanagan NS, Gonzalez C, Peterson E, Jiggins CD, Johnston JS, Heckel DG, McMillan WO. 2005. First-generation linkage map of the warningly colored butterfly *Heliconius erato*. *Heredity* 94: 408 – 417.

Wallberg A, Glémin S, Webster MT. 2015. Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera*. *PLOS Genetics* 11: e1005189. doi: 10.1371/journal.pgen.1005189

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Research* 27: 757-767.

Witherspoon DJ, Watkins WS, Zhang Y, Xin J, Tolpinrud WL, Hedges DJ, Batzer MA, Borde LB. 2009. Alu repeats increase local recombination rates. *BMC Genomics* 10: 530.

Wright SI, Agrawal N, Bureau TE. 2003. Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. *Genome Research* 13: 1897 – 1903.

Xu YX, Du JC. 2014. Young but not relatively old retrotransposons are preferentially located in gene-rich euchromatic regions in tomato (*Solanum lycopersicum*) plants. *The Plant Journal* 80: 582 – 591.

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