RESEARCH

Linkage mapping and genome annotation give novel insights into gene family expansions and regional recombination rate variation in the painted lady (*Vanessa cardui*) butterfly

Daria Shipilina[1,2*†], Karin Näsvall[1†], Lars Höök[1], Roger Vila[3], Gerard Talavera[4] and Niclas Backström[1]

Abstract

Gene family expansions and crossing over are two main mechanisms for the generation of novel genetic variants that can be picked up by natural selection. Here, we developed a high-density, pedigree-based linkage map of the painted lady butterfly (*Vanessa cardui*) - a non-diapausing, highly polyphagous species famous for its long-distance migratory behavior. We also performed detailed annotations of genes and interspersed repetitive elements for a previously developed genome assembly, characterized species-specific gene family expansions and the relationship between recombination rate variation and genomic features. Identified expanded gene families consisted of clusters of tandem duplications with functions associated with protein and fat metabolism, detoxification, and defense against infection - key functions for the painted lady's unique lifestyle. The detailed assessment of recombination rate variation demonstrated a negative association between recombination rate and chromosome size. Moreover, the recombination landscape along the holocentric chromosomes was bimodal. The regional recombination rate was positively associated with the proportion of short interspersed elements (SINEs), but not the other repeat classes, potentially a consequence of SINEs hijacking the recombination machinery for proliferation. The detailed genetic map developed here will contribute to the understanding of the mechanisms and evolutionary consequences of recombination rate variation in Lepidoptera in general. We conclude that the structure of the painted lady genome has been shaped by a complex interplay between recombination, gene duplications and TE-activity and that specific gene family expansions have been key for the evolution of long-distance migration and the ability to utilize a wide range of host plants.

Keywords: genomics; recombination; linkage map; gene family; painted lady; Lepidoptera

Introduction

The genomic era opens up opportunities for investigating relationships between genotypes and complex phenotypes on a novel level and for a better understanding of genome evolution. Combinations of different approaches can lead to novel insights into the dynamics of recurring duplications, deletions and other types of structural rearrangements, for example, by assessing molecular mechanisms and evolutionary consequences of gene family expansions and contractions, the activity of selfish genetic elements (e.g. transposable elements, TEs) and recombination rate variation.

Gene duplication has since long been recognised as an important mechanism for generating novel genetic material for natural selection to act upon [1, 2, 3], and gene family expansions and contractions are important sources for generation of phenotypic diversity [4, 5]. Comparative approaches, such as orthology analysis, allow for identification of expanding or contracting gene families and annotation of orthogroups with the functional relevance in the evolution of lineage-specific traits. This approach might be beneficial for investigating complex phenotypes such as migratory behavior or polyphagy, where combined effects of different types of genetic changes likely underlie the trait [6].

Since the spearheading work by McClintock [7], transposable elements (TEs) have been acknowledged as major contributors to different types of evolutionary change in eukaryotes [8, 9]. Transposable elements are capable of self-replication within the host genome and can mediate both small scale deletions and duplications, large scale chromosomal rearrangements [9] and considerable genome size expansions [10, 11]. In

*Correspondence: daria.shipilina@ebc.uu.se
[1]Evolutionary Biology Program, Department of Ecology and Genetics, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden
[2] Swedish Collegium for Advanced Study, Thunbergsvägen 2, 75236, Uppsala, Sweden
Full list of author information is available at the end of the article
†authors contributed equally

addition, TE insertions can affect gene function when regulatory or coding regions are targeted. Therefore, characterisation of the TE repertoire is key to understanding the microevolutionary dynamics within the genome of a species and the potential effects of TE activity on trait variation within populations and between species.

Besides gene duplication and TE activity, recombination is a process crucial for evolutionary innovation. Meiotic recombination shuffles existing segregating genetic variants, resulting in the generation of novel haplotypes [12]. Recombination also influences selection efficiency by directly preventing the accumulation of deleterious alleles (Müller's ratchet) and breaking the physical linkage between mutations with different selective effects (Hill-Robertson effects). The rate of recombination can vary on different scales. Of particular interest for population genetic processes is the variation in recombination rate between different genomic regions. Such spatial variation in the recombination rate has been observed in many different organisms [13, 14], but might be of a special interest in butterflies, which have holocentric chromosomes. Some of the important features of the latter are that the meiotic spindle attachments are not restricted to a concentrated centromere, but distributed along the chromosome. Holocentricity could hence lead to a more even recombination landscape where restrictions to the numbers and positions of crossover events are mainly limited by chiasma interference [15]. However, besides detailed recombination maps in the butterfly genus *Heliconius* [16], little is known about how the rate of recombination rate varies across chromosome regions in Lepidoptera and how recombination is associated with different genomic features [17, 18].

As indicated above, incorporating different approaches is essential for studying the genetic underpinnings of complex phenotypes and the mechanisms governing microevolutionary processes. The painted lady, *Vanessa cardui*, represents a key study system for a wide array of evolutionary studies. It is the most widespread of all butterfly species [19], and its migratory behavior includes a diverse repertoire of distinct phenotypes. In general, migratory butterfly species need to sustain long-distance flight and have well developed navigational abilities [20, **?**]. Therefore, traits related to energy metabolism, sensory reception and the flight machinery have likely been under strong directional selection. In contrast to many other migratory butterflies inhabiting temperate zones, the painted lady is a non-diapausing, multigenerational migrant, with an annual migratory circuit covering areas with extreme environmental heterogeneity [21, 22]. Despite the high risks associated with such a migratory lifestyle, painted ladies have successfully colonized almost all continents, and the species harbors high levels of genetic diversity, indicating a large effective population size [23]. This could also be a consequence of the species' ability to utilize a wide range of host plants [24, 25]. Until the era of high-throughput sequencing, the possibilities to gain insights into how the migratory and generalist lifestyle has been manifested at the level of the genome have been limited: genetic basis of migratory behavior in insects has only been investigated in a few model species (e.g. the monarch butterfly, *Danaus plexippus*) so far [26].

A key step for genomic analyses is the development of a high-contiguity genome assembly of the focal species and a thorough genome annotation. A powerful method to ensure the spatial correctness of a chromosome level physical assembly is construction of a linkage map. In this study, we present the first detailed linkage map of the painted lady and verify scaffolds from a previously available genome assembly based on long-read sequencing technology [27]. We use the genome annotation and linkage information to quantify lineage-specific patterns of gene family evolution, relative TE abundance and how the regional recombination rate variation is associated with genomic features in the painted lady. Our analyses complement earlier efforts to establish genomic tools for this species [28, 29] and give novel insights into the overall genome structure, recombination rate variation and lineage-specific gene family expansions in this species, information that informs on the molecular mechanisms underlying genome evolution in butterflies in general and the formation of the complex migratory phenotype and generalist lifestyle of the painted lady in particular.

Results

Linkage map and genome annotation

To verify a chromosome level assembly of the painted lady [27] and to get access to detailed recombination rate data, we constructed a pedigree-based linkage map. The total distance of the linkage map was 1,516 centiMorgan (cM) and contained 1,323 markers. When anchored on the 424 Mb physical assembly, the average marker density was 3.09 markers / Mb. The genome assembly was highly collinear with the marker order in the linkage map (Pearson's correlation coefficient; $R = 0.91\text{-}1.00$, p-value $> 1.00\times10^{-04}$, Figure S1) and consisted of 30 autosomes and the sex chromosomes Z and W. The high collinearity between linkage groups and assembled scaffolds, the large scaffold N50 (14.6 Mb) and high BUSCO scores (97% complete arthropod genes) confirm that the scaffolds in the assembly essentially represent complete chromosomes that could

be used for accurate characterization of genomic features and quantification of regional recombination rate estimates.

In total, TEs constituted > 150 Mb (37.40%) of the assembly and LINEs and SINE were the most abundant of the characterized repeat classes (Table 1). After automatic annotation and subsequent manual curation, 13,161 protein-coding genes were identified (including 89.90% BUSCO genes), of which 12,209 had functional annotation information (Table 1). Visual inspection of the spatial distribution of genes and TEs along chromosomes revealed rather similar distributions of repeat classes between autosomes and the Z-chromosome, but also an observable excess of repeats on smaller autosomes and a striking difference in repeat composition and gene density on the W-chromosome (Figure 1).

Table 1 Linkage map, genome assembly and annotation statistics

Linkage map	
Total map length (cM)	1,516
Number of markers	1,323
Markers per physical distance (N / Mb)	3.09
Genome assembly	
Scaffold N50 (bp)	14,615,999
GC content	33.41%
Total repeat proportion	37.40%
Repeat content (% of total repeat proportion)	
SINEs	7.30%
LINEs	14.94%
LTR elements	2.47%
DNA elements	3.04%
Simple and unknown repeats	30.17%
Gene annotation	
BUSCO genes	89.90%
Number of protein coding genes	13,161
Number of genes with functional annotation	12,209

Synteny

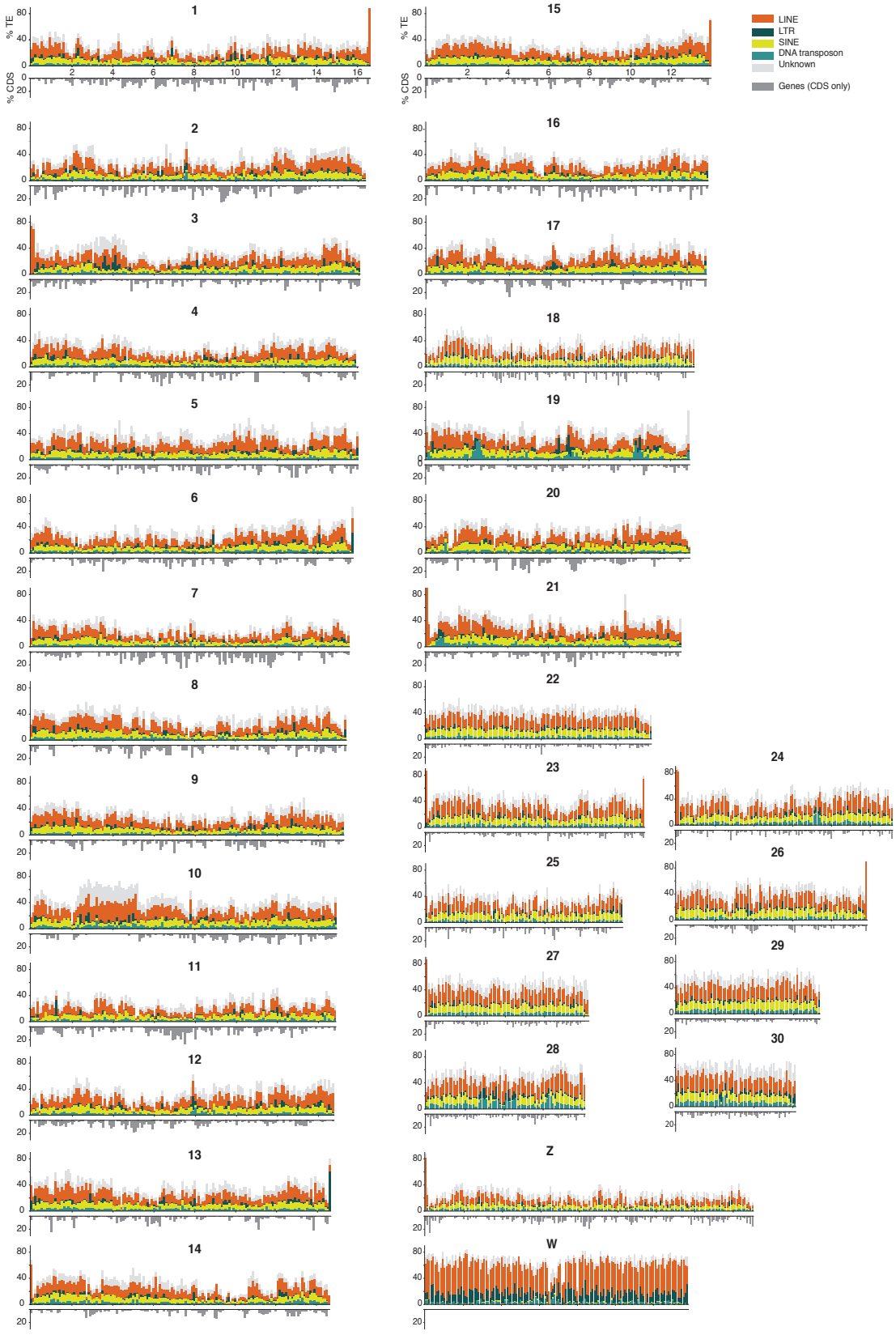
The level of large-scale structural conservation of the painted lady genome was assessed by comparing gene order on the painted lady chromosomes to two previously available high-contiguity lepidopteran genome assemblies positioned at different levels of divergence in the lepidopteran tree of life, the silkmoth (*Bombyx mori*) and the postman butterfly (*Heliconius melpomene*). Overall, the synteny was highly conserved between the painted lady and the other species, but chromosomes 28 and 26 mapped to the same chromosome (24) in *B. mori* and the previously described fusions of several chromosomes in the *H. melpomene* genome [30] could also be verified (Figure 2). In summary, this confirms that the painted lady karyotype is highly similar to the inferred ancestral butterfly karyotype [31].

Gene family evolution

To investigate the turnover of specific gene families in the painted lady, we analyzed a set of nine representative nymphalid species with detailed annotation information (see methods). The non-migratory Kamehameha butterfly (*Vanessa tameamea*) was included to assess differences in gene family evolution between sedentary and migratory lineages within the *Vanessa* genus. We found that 93.2% (1,288,332) of the total number of genes from the nine nymphalid species were clustered in 14,027 orthogroups. The percentage of genes assigned to orthogroups varied from 86.7 to 99.6% in the different species (Table S1). In the painted lady, 96.4% (12,692) of the annotated genes were assigned to 10,361 orthogroups with 19 lineage-specific orthogroups containing 63 genes (Table S1). Within the *Vanessa* genus, 65 expansions had occurred on the ancestral branch, 648 on the *V. cardui* branch and 1,563 on the *V. tameamea* branch.

We used a maximum likelihood model to detect genes with distinct gene family expansion rates in the painted lady compared to the other species. The analysis showed that 12 orthogroups were significantly expanded in the painted lady. These orthogroups contained 77 genes, of which 34 had associated GO-terms (Figure 3). Among the largest expanded gene families were two classes of proteases, a lipoprotein receptor and the Lepidoptera-specific moricin immune-gene family. Analysis of the spatial distribution of extended orthogroups revealed clustering/tandem duplications for all except one of the orthogroups (Figure 3C). Significantly enriched GO-terms for expanded gene families in the painted lady were predominantly associated with protein degradation, muscle function and development, and fatty acid and energy metabolism (Figure 3A). Multiple ontology terms were shared between expanded orthogroups, pointing towards similar functions associated with the different gene families (Figure 3B). Additionally, we identified gene families with a distinct gene expansion rate in both the painted lady and the monarch butterfly *Danaus plexippus* - the latter a key model organism for insect migration studies - and compared those to the other nymphalids. This analysis revealed 11 orthogroups with a higher expansion rate and 29 orthogroups with genes specific to these two lineages. The common orthogroups included 112 genes and were significantly enriched for GO-terms predominantly associated with metabolic processes, defence against infection and neuronal activity (Figure S2).

Figure 1 Distribution of repeat classes and genes as estimated along the painted lady chromosomes (100 kb windows). Density (% of the window covered) of different TE classes are illustrated with distinct colors cumulatively added on top of each other above the X-axis and density of genes below the X-axis (legend to the top right).



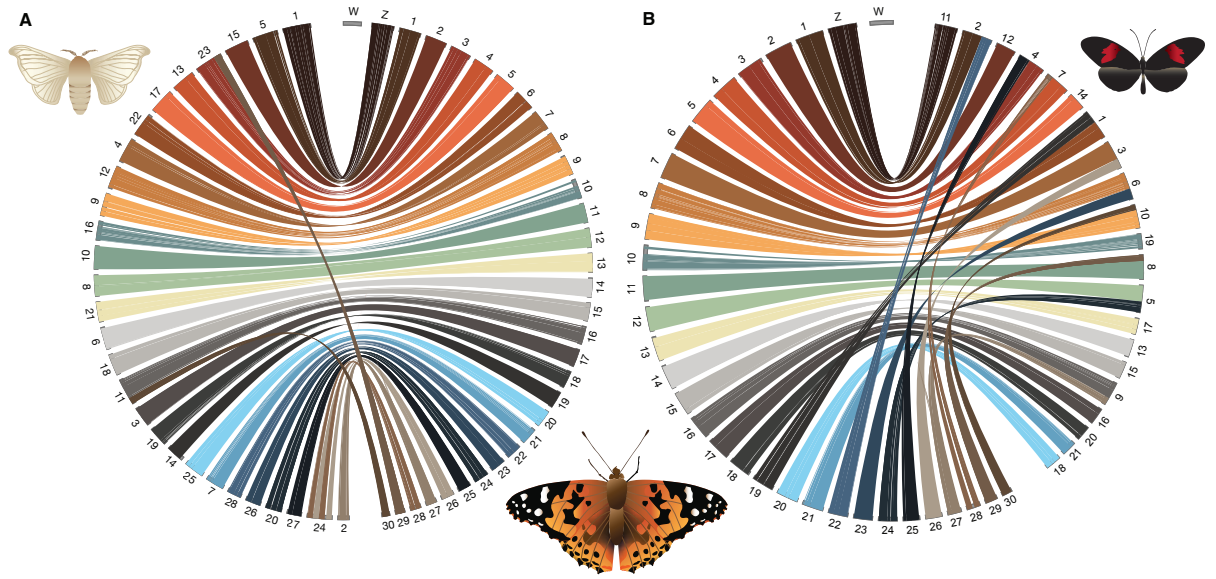


Figure 2 Synteny between the painted lady and a) the silkmoth (*Bombyx mori*) and b) the postman butterfly (*Heliconius melpomene*) chromosomes, respectively. The painted lady autosomes are sorted and named according to length and the sex chromosomes are indicated with Z and W. Orthologous genes are connected with lines. Colors represent individual chromosomes.

Patterns of recombination rate variation

Global and chromosome specific recombination rates

The development of a detailed linkage map allowed both for estimating the global recombination rate in the painted lady and to investigate potential regional recombination rate variation and association with genomic features. The average, genome-wide recombination rate was 3.81 cM / Mb (W-chromosome excluded), but there was considerable inter-chromosomal variation (2.21 - 8.00 cM / Mb; Table S2, Figure S3), with a significantly higher rate on shorter chromosomes than on longer chromosomes (Figure 4). The recombination rate on the Z-chromosome was 3.09 cM / Mb, lower than the average unweighted autosomal rate. However, the recombination rate on the Z-chromosome was not lower than expected given the overall negative correlation between recombination rate and chromosome size (Spearman's rank correlation, $\rho = $ -0.83, p-value $= 6.51 \times 10^{-07}$; Figure 4A). Besides the negative association between chromosome size and recombination rate, we also found significant negative associations between chromosome size and GC-content ($\rho = $ -0.65, p-value $= 8.35 \times 10^{-05}$) and repeat density ($\rho = 0.77$, p-value $= 1.37 \times 10^{-06}$), and a positive association with gene density ($\rho = 0.68$, p-value $= 2.63 \times 10^{-05}$)(Figure 4B-D).

Intra-chromosomal variation in recombination rate and associations with genomic elements

To quantify potential regional variation in recombination rate within chromosomes, we estimated the recombination rate in 2 Mb non-overlapping windows along each individual chromosome. The average rate across windows was similar to both the global rate estimate across chromosomes (4.05 +/- 2.45 cM / Mb) and the overall chromosome level estimates (2.58 - 7.53 cM / Mb, W-chromosome excluded). The recombination rate estimates for individual windows ranged between 0 - 14.79 cM / Mb (Figure S3, Table S2) and visual inspection revealed a bi-modal distribution with reduced recombination rate in the center of chromosomes and towards chromosome ends (Figure 4 E-H). To test this observation formally, we analyzed the difference in recombination rate between bins representing five relative distance intervals from the center of the chromosome for all chromosomes combined and found that the recombination rate was significantly lower in the center (first bin), significantly higher in the flanking terminal (fourth) regions and then again lower at the terminal end (Wilcoxon rank sum tests, p-values = 3.70×10^{-02} - 6.30×10^{-15}; Figure S4).

To assess potential relationships between the recombination rate and genomic features in more detail, we first investigated different associations between the window-based recombination rate estimates and variation in nucleotide composition and proportions of different TEs and genes. The W-chromosome was excluded from this analysis since it is non-recombining in Lepidoptera. We found that the GC-content increased towards the ends of chromosomes and was positively associated to the regional recombination rate

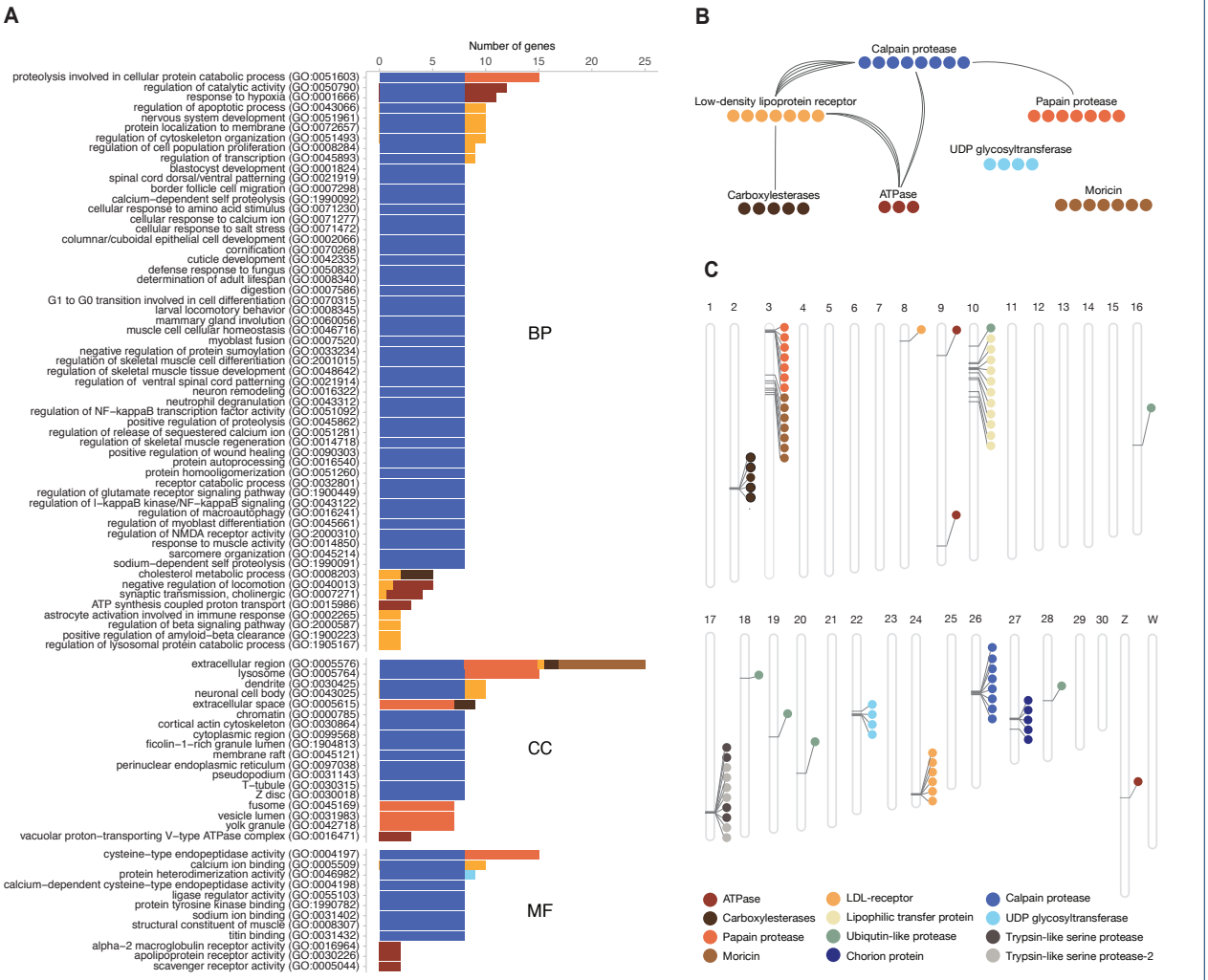


Figure 3 A) Significantly (p-value < 0.05 after FDR-correction) enriched gene ontology (GO) terms associated with expanded gene families in the painted lady. The bars show the number of genes associated with each GO-term. The different GO-categories are biological process (BP), cellular compartment (CC) and molecular function (MF). B) Orthogroups with significantly enriched GO-terms. Shared GO terms between ontology terms (biological process category only) are shown in connecting lines. C) Spatial distribution of genes from extended orthogroups identified in BadiRate analysis.

($\rho = 0.32$, p-value $= 3.68 \times 10^{-06}$). Gene density was homogeneous across chromosomes, with only a minor increase towards the chromosome center, and was negatively associated with the recombination rate ($r = -0.19$, p-value $= 7.27 \times 10^{-03}$). We found a significant positive association between the overall repeat proportion and the recombination rate ($\rho = 0.35$, p-value $= 3.48 \times 10^{-07}$; Figure 5), and this pattern was consistent for all repeat classes, but strongest for SINEs ($\rho = 0.42$, p-value $= 2.63 \times 10^{-10}$) and weakest for LTRs ($\rho = 0.14$, p-value $= 4.04 \times 10^{-02}$). The association between recombination rate and proportion of LTRs was, however, not significant when only including autosomes ($\rho = 0.11$, p-value $= 11.04 \times 10^{-02}$; Figure 5, Figure S5).

To disentangle the relative strength of associations between the regional recombination rate and genomic features, a multiple linear model was implemented with recombination rate as the dependent variable. As explanatory variables we used chromosome length, chromosome type, GC-content, proportion of genes (CDS) and proportions of all different classes of TEs. We found that the regression model was significant ($df = 197$, $F = 7.73$, p-value $= 5.36 \times 10^{-09}$) and explanatory variables in the model accounted for 21% of the variation in recombination rate ($R2 = 0.24$, $adjR2 = 0.21$). Most of the variation was explained

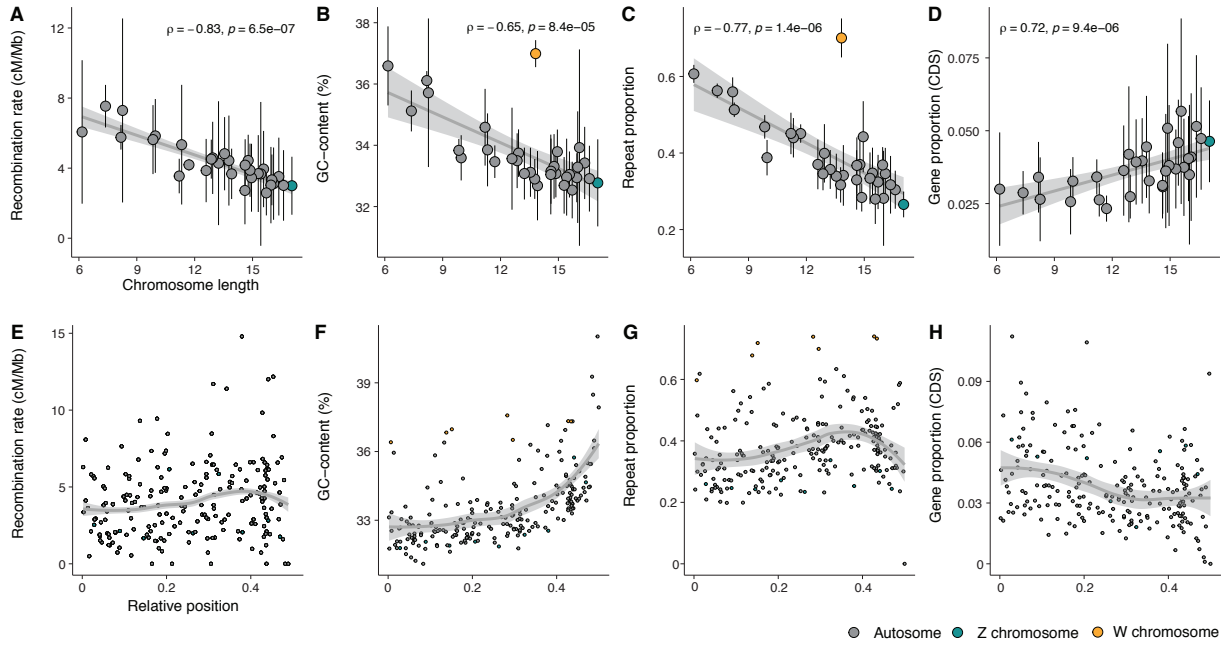


Figure 4 Top row: Associations between chromosome length and A) recombination rate, B) base composition, C) repeat and D) gene proportions. Chromosome length is given in megabases (Mb). Bottom row: Regional distribution of the recombination rate (E), base composition (F), repeat (G) and gene (H) density in 2 Mb windows along the chromosomes. All chromosomes were analyzed jointly and the x-axis shows the relative position (proportion of chromosomal length) from the center of the chromosomes.

by the positive association with the proportion of SINEs (est 1.59, p-value = 9.75×10^{-04}) and the negative association with chromosome size (est -0.48, p-value = 4.88×10^{-02}; Figure 5, Table S3).

Finally, we explored whether gene expansions could be associated with other genomic features, and we therefore compared TE abundance in the regions with and without gene gains. The mean densities of LTRs, LINEs and DNA transposons were higher in regions with gene gains (Wilcoxon rank sum test, p-value 3.1×10^{-03} - 6.0×10^{-04}; Figure S6), as was mean GC-content (p-value 3.0×10^{-02}). The gene densities or recombination rates did not differ between regions with or without gene gains (Wilcoxon rank sum tests, p-values = 9.1×10^{-01} - 8.0×10^{-01}; Figure S6).

Discussion

The genome of the painted lady butterfly

Here we present detailed results on the genomic architecture and regional recombination rate variation in the painted lady. The data paves the way for understanding the interplay between molecular mechanisms and micro-evolutionary processes shaping the genome of butterflies in general and provide the first insights into the links between genomic features and the unique lifestyle of this species. The rapid technological advances and dropping costs of DNA-sequencing methods have led to a staggering development rate of high-quality genome assemblies, including many butterfly species [32, 33, 34, 35, 36], and the availability of genomic resources will probably increase almost exponentially in the near future, as a result of the Darwin tree of Life (/https://www.darwintreeoflife.org/), the European Reference Genome Atlas (ERGA; https://www.ergabiodiversity.eu/) and other similar initiatives. However, detailed and curated genome annotation data are more time-consuming and expensive to generate and therefore still limiting comparative/population genomic and genotype-phenotype association approaches, not the least in butterflies [30, 37, 38]. Another limiting factor for understanding both genome architecture in general, the relative effects of random and selective forces on sequence evolution and maintenance/loss of genetic diversity, and divergence processes, is that detailed recombination rate data are both laborious and time-intensive to gain, especially for natural populations. As a consequence, high-density recombination maps are still lacking for the vast majority of wild species where genome assemblies are now available. The detailed annotation information and the high-density linkage map for the painted lady developed

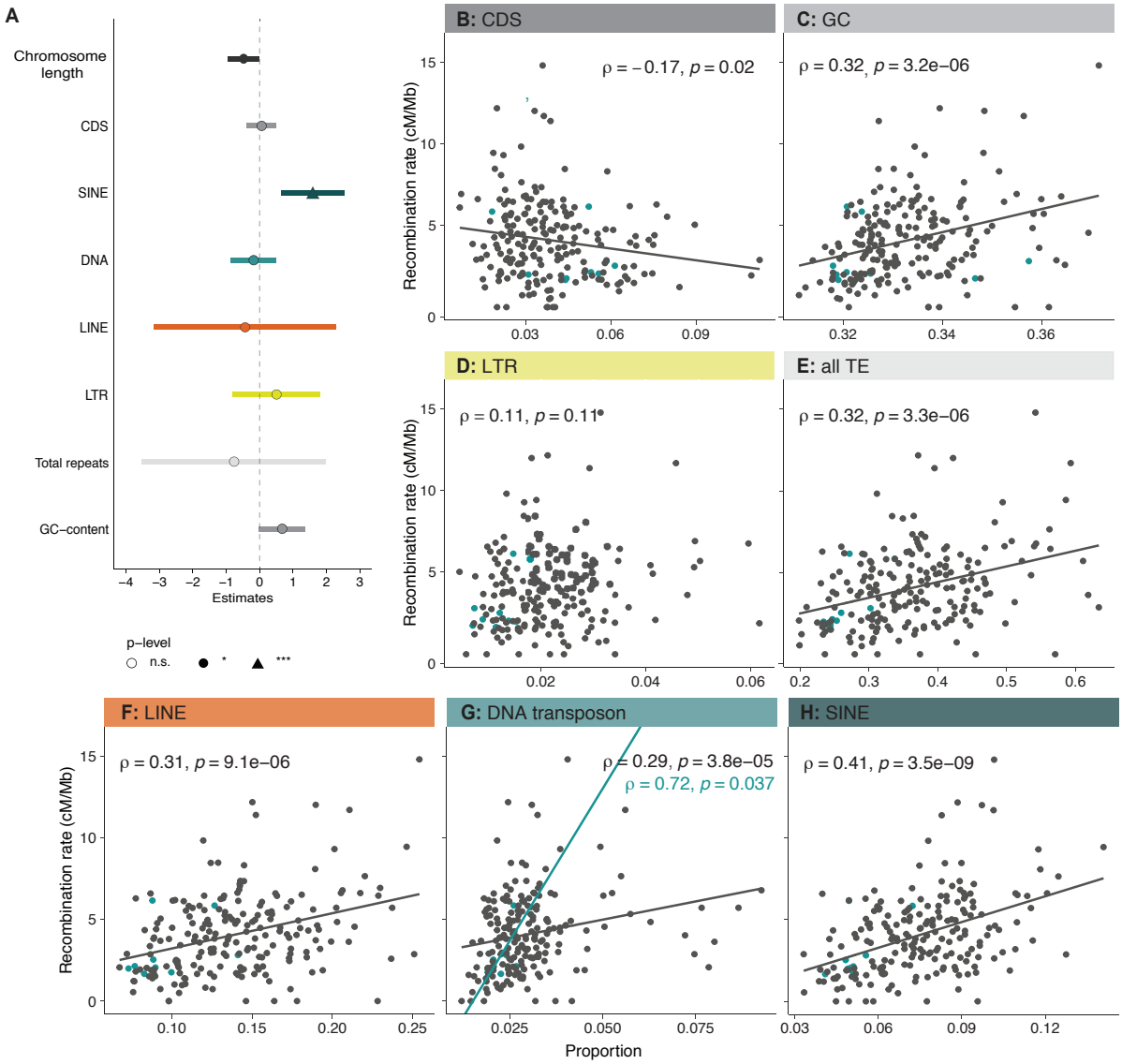


Figure 5 Correlation between recombination rate and density of genomic features. A) Summary of the linear model with regional recombination rate as the response variable. Each explanatory variable in the model is listed along the Y-axis and the relative estimated effect (X-axis), and error intervals are indicated with horizontal bars for each variable. B - H) Associations between the regional recombination rate and specific genomic features. Linear regression lines, Spearman's correlation coefficients (ρ) and the corresponding ρ-values for significant analyzes are given. Gray dots/lines indicate autosomal regions and turquoise dots/lines Z-chromosome linked regions

here, therefore provide opportunities for both comparative studies on genome structure organization, population genomic- and micro-evolutionary investigations in the entire Lepidoptera clade.

Chromosome numbers have been shown to vary considerably between different butterfly and moth species; the haploid chromosome counts range from 5 to 223 [39, 40]. In agreement with previous data [29], both the linkage map and the DToL genome assembly clearly showed that the painted lady has a total haploid chromosome count of 31. We confirmed high levels of synteny and gene order collinearity between the painted lady and the silkmoth, and the lineage specific chromosome fusions characterized before in the postman butterfly [30]. Hence, similar to other nymphalid butterflies, the painted lady has retained the inferred ancestral lepidopteran karyotype [31]. The annotation procedure revealed that the painted lady harbors a gene

set (n = 13,161) close to the suggested core set in Lepidoptera [41, 42] and a relatively low overall TE content. However, the TE content was significantly higher and the gene density lower on smaller chromosomes.

Sex chromosomes

The assembly and annotation of sex-chromosomes, especially the non-recombining parts of sex-limited chromosomes (i.e. the W-chromosome in Lepidoptera), can be technically challenging due to the high density of repetitive elements. Up to date, there are only a few Lepidoptera species where the W-chromosome has been assembled and annotated [43]. Given the high-quality assembly we had access to, we performed annotation and manual curation of TEs and coding genes for the painted lady W-chromosome. In contrast to previous annotation [27], we could not confirm the presence of any protein coding genes. Gene models created on the preliminary annotation were not confirmed after manual curation and functional domain annotation. A lack of protein coding genes on the W-chromosome has also been observed in the silkmoth [43]. This apparent complete loss of protein coding genes on the Lepidoptera W-chromosome is obviously a consequence of the degradation process that has been well described for non-recombining parts of sex-chromosomes in many systems [44]. While having a size equal to an average autosome, the W-chromosome also demonstrated a significantly higher overall proportion of TEs, a larger fraction of longer TEs, and a different distribution of repeat classes compared to other chromosomes.

Similar to the silkmoth and julia heliconian (*Dryas iulia*), the W-chromosome in the painted lady had a significantly higher proportion of LTRs and LINEs [45, 43]. The proportion of SINEs was however much smaller on the W-chromosome than on the autosomes and the Z-chromosome. A lack of protein coding genes, like we observed on the painted lady W-chromosome, has also been observed in the silkmoth [43], and is likely a consequence of the general degradation process of the non-recombining sex-chromosome [44]. The higher accumulation of TEs is also an expected consequence of recombination suppression and comparatively low effective population size (N_e) of the W-chromosome (1/4 of the autosomes at equal sex-ratios), both as a consequence of Müllers ratchet and since the overall efficiency of selection against TE insertion is reduced for non-recombining chromosomes [44]. The Z-chromosome is generally highly conserved in Lepidoptera [46] and it is the largest of all the painted lady's chromosomes. We did not find any significant differences in gene or TE content on the Z-chromosome compared to the autosomes.

Gene family analysis

Gene family expansions can provide the raw material for both neo- and sub-functionalizing evolutionary directions, and the rate of gene duplication can be significantly higher than the rate of function-altering single nucleotide mutations [47]. However, most gene duplication events are probably deleterious [48] or effectively neutral, leading to a low probability of fixation of novel gene copies [49]. We found a comparatively low proportion of lineage-specific gene duplications in the painted lady, which could be a consequence of the large N_e of the species [23], which translates to efficient selection against slightly deleterious variants. The majority of the (comparatively few) significant gene expansions in the painted lady lineage clustered on single chromosomes - only a single gene family had expanded and dispersed across multiple chromosomes - suggesting that unequal crossing over has been the main mechanism behind gene family expansions.

The painted lady has extraordinary life-history characteristics and has become a quickly uprising complementary model organism for studying insect migration. Over most of the almost cosmopolitan distribution range [50], the painted lady individuals annually complete a multigenerational migratory circuit that can span many thousand kilometers in total, and single individuals can migrate $> 4,000$ kilometers during lifetime [21]. In addition, the painted lady is a polyphagous generalist that utilizes > 300 different larval host-plants in 11 plant families [24, 51, 52] and, in contrast to other migratory butterflies like the monarch and the red admiral (*Vanessa atalanta*), the painted lady is non-diapausing [50]. This unique combination of life-history traits is accompanied by high levels of heterozygosity and presumably a very large effective population size [23]. The genetic underpinnings of migratory behavior have only been preliminarily characterized for a handful of insect species [53, 54] and have not been studied in painted lady before. The dissection of potential associations between genetic (and epigenetic) variants and complex phenotypes like 'migratory behavior' obviously requires a combination of multiple approaches. As the first step to understanding lineage-specific characteristics of the painted lady, we here focused on gene family evolution. Our results showed a limited number of genes with significant copy number expansions unique to the painted lady lineage. The expanded gene families were mainly associated with functions related to the transport of fatty acids, protein metabolism, and muscle structure and activity. Since migratory insects mainly use fat as an energy resource during migration [55, 56, 57, 58], both the capacity to build up fat deposits and efficient sequestration of fatty acids have likely been un-

der strong selection in the painted lady. Likewise, enhanced muscle structure and function should be advantageous for long-distance migrants compared to sedentary species. Therefore, efficient fine-tuning and optimization of fatty acid metabolism and increased muscle sustainability during migration could have been aided by the expansion of specific gene sets involved in those processes.

Besides the obvious advantages of having efficient energy metabolism and high-functioning flight machinery, long-range migrants will also benefit from utilizing a multitude of different host plants since they will encounter dramatically different habitats, both during the lifespan of single migratory individuals and between consecutive generations. In contrast to the monophagous monarch butterfly, the painted lady can utilize a wide range of host plant species [52, 25], an adaptation that probably has been coupled to strong selection on genes involved in detoxification of secondary metabolites. We found that two of the significantly expanded gene families in the painted lady (UDP-glycosyltransferase, carboxylesterase) were associated with detoxification and polyphagy [59, 60, 61]. The UDP-glycosyltransferase superfamily includes Lepidoptera-specific subfamilies associated with a variety of functions, such as affinity for plant secondary metabolites [62, 63]. In the painted lady larvae, it is upregulated in response to utilization of an extended range of hostplants [24]. Copy-number expansions of these detoxifying gene families could have allowed the painted lady to increase the range of host plants that can be utilized and consequently paved the way for developing the non-diapausing, multigenerational, long-distance migratory lifestyle. The wide range of habitats that long-distance migratory species encounter also probably means that they are exposed to many more different pathogens than sedentary species. Our analysis revealed that the Lepidoptera-specific gene *moricin*, associated with inducible antimicrobial peptides [64], was significantly expanded in the painted lady. An increase in the number of *moricin* copies could have increased the efficiency of defense against a larger suite of pathogens.

The genetic basis of migratory behavior has been investigated in some detail in the monarch butterfly. A combination of approaches has identified candidate genes associated with orientation, chemoreception and regulation of the circadian clock [65, 54]. Migratory behavior has evolved independently multiple times within the Papilionoidea clade [66] and in the *Vanessa* genus [67], and the life histories of the monarch butterfly and the painted lady are distinct. However, long-distance migration should put selective pressure on similar traits (e.g. navigation, energy metabolism, muscle endurance), and it is therefore possible that specific gene categories have been under selection in independent lineages. We therefore expanded the analysis to include gene-family expansions that were shared between the painted lady and the monarch in our sample set. Significantly expanded gene families were enriched for functions associated with various metabolic processes, defense against pathogens and neuronal activity, all of which are straightforward to intuitively associate with migratory behavior. One gene family with expanded copy numbers in both species and an especially pronounced increase in the painted lady was a family of vacuolar ATPases. The ATPases are ATP-dependent proton pumps involved in membrane transports, and they have been shown to affect for example ion transport in insects [68]. Given the unique expansion of this gene family in both species, we speculate that copy number increase could be involved in flight muscle coordination and/or ion transport for maintenance of homeostasis during long periods of flight.

In this study, we get a first glimpse of the specific genes that have undergone copy number expansions in painted lady specifically and independently in the two migratory species. The functions associated with the expanded gene families can be coupled to the evolution of long-distance migratory behavior. However, further studies on larger species sets with independent migratory and sedentary sister species pairs, in combination with detailed intraspecific population genetic analysis and functional verification experiments will be necessary to dissect the genetic underpinnings of migratory behavior in butterflies in detail.

Patterns of recombination rate variation

Detailed data on recombination rate variation are crucial for understanding the relative effects of random genetic drift and selection on levels of genetic diversity and disentangling the evolutionary forces shaping genetic divergence between incipient species. Understanding how recombination breaks down linkage disequilibrium between physically linked regions is also important for the efficient design of association studies aimed at coupling genetic variation to phenotypic traits. Despite these important contributions to evolutionary genomics research, detailed recombination maps are only available for a handful of butterfly species [69, 32, 30, 70, 35, 71]. In some cases, linkage maps have been used to improve and/or verify the correctness of physical genome assemblies, but analysis of the recombination rate has not been thoroughly assessed in many butterfly species. Here we developed a high-density linkage map based on segregation information in a pedigree with 95 offspring. The map contained $> 1,300$ ordered markers and the overall

density was > 3 markers per Mb. Despite being based on a single pedigree, the genetic map developed here revealed a recombination landscape in strong agreement with what has been observed in other butterflies [30, 16]. This indicates that the painted lady genetic map reflects the historical recombination landscape in the species well.

We estimated the genome-wide average recombination rate in the painted lady to be 3.81 - 4.05 cM / Mb, dependent on the method applied. The global rate was in the lower end of recombination rate estimates from other Lepidoptera species, which have been in the range from 2.97 - 4.0 cM / Mb in the silkmoth [72, 73] to 5.5 - 6.0 cM / Mb in different *Heliconius* species [74, 75]. We found a significant negative association between chromosome length and the recombination rate in the painted lady. This is a consistent pattern found across many organism groups and likely a consequence of that at least one crossover event is necessary for correct segregation of chromosomes during meiotic division in the recombining sex, leading to a higher recombination rate per unit length for shorter chromosomes [17, 76, 16]. Butterflies and moths are holocentric, i.e. they lack distinct centromere regions which means that the spindle fibers can attach 'anywhere' along the chromosomes during cell division. This might lead to an expectation of a more uniform distribution of recombination events along chromosomes in holocentric species if crossovers occur randomly. The window-based analysis in the painted lady revealed a bimodal distribution of recombination events along chromosomes, with a significantly higher rate in regions close to, but not directly at, chromosome ends. This distribution is in agreement with previous observations, both in Lepidoptera and in other animals with different centromere types [17, 16]. A possible explanation for this pattern could be mechanical or tension interference between chiasmata when > 1 recombination event occurs on the same chromosome during the same meiotic division [17]. However, in the holocentric *Caenorhabditis elegans*, the number of recombination events is limited to precisely one per chromosome per meiosis, but there is still a strong bimodal pattern of recombination rate variation along chromosomes in this species [15]. An alternative explanation for the bimodal distribution of recombination events along chromosomes could be that synaptonemal complexes are directed towards specific physical positions when the telomeres attach to the nuclear wall [77]. As indicated above, we also found that the recombination rate dropped significantly at the far ends of the chromosomes in the painted lady. This reduced recombination rate at chromosome ends is also consistent with earlier observations and could potentially be attributed to selection against synaptonemal complex formation at chromosome ends, due to a higher risk of ectopic recombination in these generally repeat-rich regions [78].

Since recombination is directly associated with the efficacy of selection, a negative correlation between the regional recombination rate and a number of repeats would be expected if TE insertions predominantly are deleterious. Such associations have been observed in many organisms, although the relationship between TE-abundance and the recombination rate varies to some extent across species and different TE-classes [79, 80]. In the painted lady, we observed a significant positive association between TE-abundance and the regional recombination rate, predominantly driven by a strong effect of SINE density. An explanation for the strong association between SINE density and recombination rate could be SINE-mediated recombination, as has for example been described in humans [81]. In addition to the strong positive association between SINE density and recombination rate on the autosomes and the Z-chromosome, the absence of SINEs on the non-recombining W-chromosome supports that SINEs might be able to hijack the recombination machinery. However, we can not exclude other factors affecting both the recombination rate and the proliferation efficiency of SINEs. For example, both synaptonemal complexes and SINE insertions might be directed towards regions of more open chromatin structure.

In contrast with results from similar studies in other organism groups [82, 83], we observed a negative association between the recombination rate and gene density. This is likely a consequence of the strong association between recombination rate and chromosome size, since the association with gene density was insignificant when chromosome size was included as an explanatory variable. The observed weak positive association between GC-content and recombination is in agreement with the limited effect of GC-biased gene conversion (gBGC) in butterflies [84]. We did not find any association between recombination rate and the presence of extended orthogroups, which would be expected if gene duplication is associated with unequal crossing-over. This could possibly be a consequence of the more efficient removal of deleterious duplications in regions with higher recombination rate. However, repetitive elements can trigger ectopic recombination which can explain the observed significant positive association between gene gains and density of LTRs, LINEs and DNA elements in the painted lady.

Conclusions

In this study, we present detailed annotation and recombination rate information for the painted lady butterfly (*Vanessa cardui*), a species with a remarkable

life-history traits such as long distance migration, continuous direct development and a capacity to utilize many different types of larval host plants. We analyzed lineage-specific gene family expansions and found that expanded genes were mainly associated with fat and protein metabolism, detoxification and defense against pathogens. A detailed TE-annotation revealed that several TE-classes were positively associated with the presence of gained genes, potentially indicating their involvement in ectopic recombination. Recombination rate variation was negatively associated with chromosome size and positively associated with the proportion of short interspersed elements (SINEs). We conclude that the genome structure of the painted lady has been shaped by a complex interplay between recombination, gene duplications and repeat activity and provide the first set of candidate genes potentially involved in the evolution of migratory behavior in this almost cosmopolitan butterfly species.

Methods

Linkage map

Sampling and DNA-extraction

Offspring from one painted lady female were reared on thistles (*Cirsium vulgare*) in the greenhouse until pupation. The bursa copulatrix of a female was examined and only one spermatophore was detected, indicating that a single male had sired all offspring. The offspring were snap frozen in liquid nitrogen and stored in −20°C until DNA extraction. DNA was extracted from thorax tissue of the female and an abdominal segment of the offspring pupae, using a modified high salt extraction method [85]. The quality of the DNA was analyzed with Nanodrop (ThermoFischer Scientific) and the yield was quantified with Qubit (ThermoFischer Scientific). Extracted DNA was digested with the restriction enzyme EcoR1 according to the manufacturer's protocol, using 16 hours digestion time (ThermoFischer Scientific). DNA fragmentation was verified with standard gel electrophoresis. Digested DNA from 95 offspring with the highest yield and the dam was shipped to the National Genomics Infrastructure (NGI, see acknowledgements) in Stockholm for library preparation (standard protocol), individual barcoding and multiplex sequencing using 2×151 bp paired-end reads on one NovaSeq6000 S4 lane.

Building the linkage map

The quality of the raw reads was assessed with FastQC [86]. The reads were filtered using the Stacks2 modules `clone_filter` to remove PCR-duplicates and `process_radtags` to remove reads if the average phredscore dropped below 10 in any window 15% of the length of the read [87]. Removal of reads with unassigned bases and truncation to 125 bp was done using option `-c`, and `--disable_rad_chec` was applied to keep reads with incomplete RAD-tags.

We mapped the filtered reads to the previously published genome assembly [27] using the bwa mem algorithm [88] with default options. Resulting bam files were sorted with samtools sort [89] and filtered with `samtools view –q 10` (only reads with mapping quality score above 10 were retained). A custom script was applied to retain reads with unique hits only. The mapping coverage was analyzed with Qualimap [90]. The offspring were defined as females if the coverage on the Z-chromosome was $< 75\%$ of the average coverage over all chromosomes and as males if the coverage was $> 75\%$. Samtools mpileup was used for variant calling using minimum mapping quality (`-q`) 10 and minimum base quality (`-Q`) 10 [89]. The variants were then converted to likelihoods with Pileup2Likelihoods in LepMap3 using default settings [91]. The LepMap3 protocol [91] with some modifications was used to construct the linkage map (Supplementary methods 1).

Genome annotation and whole genome statistics

Genome assembly statistics

With very few exceptions, the order of markers in the linkage map was in agreement with the physical order in the assembly. We therefore did not make any corrections to the physical assembly before further analysis. Standard genome assembly summary statistics were calculated for the genome assembly after linkage map verification, using the QUAST suite [92]. For the subsequent analysis we excluded unassembled haplotigs from the genome assembly and retained all the other scaffolds.

We used MCScanX [93] to detect syntenic blocks between the painted lady genome assembly on the one hand and the silkmoth and the the postman butterfly on the other. We downloaded the annotation for the silkmoth assembly from SilkBase (/https://silkdb.bioinfotoolkits.net/) and the 2.5-version of the postman annotation from LepBase (download.lepbase.org/v4/, downloaded 2021-06-21). BLAST was used for primary alignment and we used a custom script to select the five hits with the highest E-values and used them as input for the MCScanX. The CIRCOS library [94] was used for visualization of the results.

Gene and repeat annotation

The annotation of the painted lady genome assembly was performed using MAKER version 3.00.0 [95] iteratively in three steps. In the first step, we mapped previously available transcriptomic evidence data from the painted lady based on wing tissue [28](accessed on 2020-05-15) and masked all known repeats. RepeatMasker version 4.0.3 [96] was used within the MAKER

pipeline with a manually curated Lepidoptera repeat library [11] serving as a reference. The first MAKER round produced a set of gene models, which were quality controlled using Annotation Edit Distance (AED) statistics. AED quantifies congruency between a gene annotation and its supporting evidence. We discarded gene models with AED scores higher than 0.5 (50% of the gene model length not matching the corresponding evidence sequence) using custom scripts. The retained gene models were provided as a training set for the second run of MAKER.

The second iteration of MAKER was run to generate gene models using the *ab initio* gene predicting algorithm implemented in SNAP [97]. For the last step in the MAKER pipeline, gene models predicted by SNAP and additional protein evidence from the Uniprot database (https://www.uniprot.org/; accessed 2021-04-01) were used. A set of Lepidoptera proteins from the Swiss-prot section of the Uniprot database were downloaded and manually curated. All genes from the curated set were included while only fully sequenced nuclear proteins with predicted functions from the non-curated gene set were included (custom scripts were used for selection). This selection resulted in 36,907 proteins. Finally, all obtained evidence and *ab initio* predicted genes were merged resulting in 18,860 gene models. Resulting genes were renamed using MAKER supplementary scripts.

Manual curation

The gene models constructed by MAKER were filtered based on standard options; discarding gene models with AED and eAED scores < 0.5 and/or length < 50 amino acids. To search for functional domains in putative genes, we used InterProScan with default settings [98]. A number of TE-related domains were detected within gene models indicating a need for a more detailed transposable element annotation. The repeat library was extended by adding evidence from the current RepBase for all Arthoropoda [99] and curated repeats from the monarch butterfly [100]. RepeatModeller [] and RepeatMasker [95] were thereafter run again with the updated repeat database.

Coordinates of newly identified repeats were intersected with gene model positions using BEDTools [101] and we removed gene models that overlapped more than 50% of the length of a repeat. We then searched for keywords in the InterProScan domain output and removed genes containing at least one TE domain. For the W-chromosome, we manually curated the InterProScan output and found no functional information for any genes. The filtering resulted in a total set of 13,161 genes, including 12,209 genes with preliminary assignments in eggnog [102]. The entire painted lady gene set represented 89.9% of the complete BUSCO arthropod gene set [103].

Gene family evolution

We investigated gene family evolution in the painted lady by comparing our obtained gene annotations with other annotated nymphalid genomes available on Lepbase. Protein fasta files for eight other nymphalid species - squinting bush brown (*Bicyclus anynana*), monarch (*Danaus plexippus*), postman (*Heliconius melpomene melpomene*), red postman (*Heliconius erato lativitta*), common buckeye (*Junonia coenia*), ringlet (*Maniola hyperantus*), speckled wood (*Pararge aegeria*) and Kamehameha butterfly *(Vanessa tameamea)* - were downloaded from Lepbase (http://download.lepbase.org/v4/sequence/: accessed 2021-06-21; Supplementary Table S1). The annotated gene sets in each species were filtered to only include one transcript per gene. To cluster the annotated genes into orthogroups and infer species specific orthogroups and gene duplications, OrthoFinder/2.5.2 [104] was used with default settings. The total gene counts for each orthogroup and species from OrthoFinder was used as input to estimate gene family expansions and contractions with the software BadiRate, using the maximum likelihood option and the birth/death/innovation (BDI) model [105]. We used the species tree obtained with OrthoFinder as input, with additional conversions using Tree as implemented in ete3 [106].

For each orthogroup identified in OrthoFinder, different models reflecting the evolution of the gene families were tested. The null model (Global rate model) assumes a uniform rate of gene gains/losses for all branches in the provided species tree. Alternative models were specified as follows; i) to detect gene family changes specific to the painted lady, a distinct branch rate was specified in the painted lady with all the other branches evolving at uniform, background rates, and, ii) the terminal branches of the two distinct migratory species, the painted lady and the monarch, were set at a common rate, differing from all other taxa in the data set. The rationale behind the last setting was to allow for identification of gene family expansions shared between the painted lady and monarch butterfly. Each model was run twice and the replicate with highest likelihood for each model was used for model comparisons.

Likelihoods of all models were compared using Aikaike's Information Criterion (AIC) [107], calculated as $2K - 2\log L$, where K is the number of parameters and $\log L$ is the logarithm of the likelihood of the model. The orthogroups where the alternative models in BadiRate inferred gene gains > 0 with the lowest AIC, were used for analysis of functional enrichment. BadiRate was partly run with a modified version of the R-package BadiRateR (BadiRate) and custom scripts. We visualized the genomic location of genes belonging

to extended orthogroups identified in BadiRate with custom bash scripts and PhenoGram [108], using gene names and positions from the annotation gff file.

Gene ontology enrichment
Potential enrichment of functional categories in the significantly expanded gene sets was analyzed using the Bioconductor package topGO version 2.44.0 [109] in R version 4.1.0 [110]. A custom database was generated based on the annotated gene set with gene ontology (GO) terms associated to the categories biological process, cellular component and molecular function. Since the gene set of interest is based on gene counts, the enrichment test was performed with Fisher's exact test using the default algorithm (`\weight01"`) which accounts for the hierarchical structure of the GO-terms [111]. This means that the resulting tests were not completely independent and correcting for multiple testing might be over-conservative. We still adjusted the p-values with Benjamini-Hochberg's method of multiple test correction [112].

Recombination rate analysis

Chromosome level analysis
Global and chromosome-specific recombination rates were estimated by dividing the linkage map length (unit = cM) with the physical length (bp) of the corresponding part (whole genome or individual chromosome) of the physical genome assembly. Regional recombination rates were estimated with local linear regression in 2 Mb non-overlapping windows containing 2 or more markers using the R-package MareyMap [113].

Window-based analysis
We quantified the spatial distribution of various genomic features along the painted lady genome using custom scripts (available on GitHub, see Data Accessibility). Positions of the specific TE classes were accessed from the RepeatMaker output file and positions of genes were taken from the annotation file from MAKER. For each 2 Mb window, we calculated the density (fraction of window covered by element / window length) of genes, TEs in total, LTRs, SINEs, LINEs and DNA-transposons with in-house developed scripts. For the genes belonging to extended orthogroups, we integrated the list of the extended orthogroups from BadiRate, gene names from OrthoFinder and positions from MAKER. All 2 Mb windows of the genome were assigned to bins; one bin containing windows without gene gains and the other with windows containing at least one gene. Potential differences in densities of genomic features between bins were assessed with two-sided Wilcoxon rank sum tests in R [114].

To characterize the associations between recombination rate and specific genomic elements, correlation tests were performed with the cor.test function in R using Spearman's rank correlation [115], after testing for deviation from normal distribution with the Shapiro-Wilk normality test [116]. We then applied the `lm` function in R to explore the relationships between the recombination rate as response variable and chromosome length, relative position on the chromosome, density of genes, density of different repeat classes, and GC-content as explanatory variables. Prior to the latter analysis, explanatory variables were scaled and centered, by subtracting the mean and dividing by the standard deviation of the variable. We used the R-package `ggplot2` for visualizations [117].

Acknowledgements
Financial support for this project was provided by FORMAS (Research grant 2019-00670 to N.B.) and The Swedish Collegium for Advanced Science (Natural Sciences Programme, Knut and Alice Wallenberg Foundation, Postdoc funding for D.S.). R.V. was supported by the grant PID2019-107078GB-I00 funded by MCIN/AEI/10.13039/501100011033. G.T. was supported by the grant PID2020-117739GA-I00 funded by MCIN/AEI/10.13039/501100011033 and by "La Caixa" Foundation (ID 100010434) through the grant LCF/BQ/PR19/11700004. 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.

Data access
All raw data have been submitted to the European Nucleotide Archive (ENA accession number XXXX). Scripts are available in the GitHub repository (GitHub LINK XXXXX).

Competing interest statement
The authors declare no financial or other competing interests.

Author details
[1]Evolutionary Biology Program, Department of Ecology and Genetics, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden. [2] Swedish Collegium for Advanced Study, Thunbergsvägen 2, 75236, Uppsala, Sweden. [3] The Butterfly Diversity and Evolution Lab, Institut de Biologia Evolutiva, Passeig Marítim de la Barceloneta 37-49, 08003, Barcelona, Spain. [4] Institut Botànic de Barcelona (IBB, CSIC-Ajuntament de Barcelona), Passeig del Migdia s/n, 08038, Barcelona, Spain.

References
1. Henikoff, S.: Gene Families: The Taxonomy of Protein Paralogs and Chimeras. Science **278**(5338), 609–614. doi:10.1126/science.278.5338.609
2. Ojeda-López, J., Marczuk-Rojas, J.P., Polushkina, O.A., Purucker, D., Salinas, M., Carretero-Paulet, L.: Evolutionary analysis of the Moringa oleifera genome reveals a recent burst of plastid to nucleus gene duplications. Scientific Reports **10**(1), 17646. doi:10.1038/s41598-020-73937-w
3. Zhang, L.: Does Recombination Shape the Distribution and Evolution of Tandemly Arrayed Genes (TAGs) in the Arabidopsis thaliana Genome? Genome Research **13**(12), 2533–2540. doi:10.1101/gr.1318503
4. Chen, S., Krinsky, B.H., Long, M.: New genes as drivers of phenotypic evolution. Nature Reviews Genetics **14**(9), 645–660. doi:10.1038/nrg3521
5. Kondrashov, F.A.: Gene duplication as a mechanism of genomic adaptation to a changing environment. Proceedings of the Royal

Society B: Biological Sciences **279**(1749), 5048–5057. doi:10.1098/rspb.2012.1108
6. Schwander, T., Libbrecht, R., Keller, L.: Supergenes and Complex Phenotypes. Current Biology **24**(7), 288–294. doi:10.1016/j.cub.2014.01.056
7. McClintock, B.: Controlling Elements and the Gene. Cold Spring Harbor Symposia on Quantitative Biology **21**(0), 197–216. doi:10.1101/SQB.1956.021.01.017
8. Kazazian, H.H.: Mobile Elements: Drivers of Genome Evolution. Science **303**(5664), 1626–1632. doi:10.1126/science.1089670
9. Kidwell, M.G., Lisch, D.: Transposable elements as sources of variation in animals and plants. Proceedings of the National Academy of Sciences **94**(15), 7704–7711. doi:10.1073/pnas.94.15.7704
10. Podsiadlowski, L., Tunström, K., Espeland, M., Wheat, C.W.: The Genome Assembly and Annotation of the Apollo Butterfly Parnassius apollo, a Flagship Species for Conservation Biology. Genome Biology and Evolution **13**(8), 122. doi:10.1093/gbe/evab122
11. Talla, V., Suh, A., Kalsoom, F., Dincă, V., Vila, R., Friberg, M., Wiklund, C., Backström, N.: Rapid Increase in Genome Size as a Consequence of Transposable Element Hyperactivity in Wood-White (Leptidea) Butterflies. Genome Biology and Evolution **9**(10), 2491–2505. doi:10.1093/gbe/evx163
12. Peñalba, J.V., Wolf, J.B.W.: From molecules to populations: Appreciating and estimating recombination rate variation. Nature Reviews Genetics **21**(8), 476–492. doi:10.1038/s41576-020-0240-1
13. Stapley, J., Feulner, P.G.D., Johnston, S.E., Santure, A.W., Smadja, C.M.: Variation in recombination frequency and distribution across eukaryotes: Patterns and processes. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences **372**(1736), 20160455. doi:10.1098/rstb.2016.0455. 29109219
14. Tiley, G.P., Burleigh, J.G.: The relationship of recombination rate, genome structure, and patterns of molecular evolution across angiosperms. BMC Evolutionary Biology **15**(1), 194. doi:10.1186/s12862-015-0473-3
15. Barnes, T.M., Kohara, Y., Coulson, A., Hekimi, S.: Meiotic recombination, noncoding DNA and genomic organization in Caenorhabditis elegans. Genetics **141**(1), 159–179. doi:10.1093/genetics/141.1.159
16. Martin, S.H., Davey, J.W., Salazar, C., Jiggins, C.D.: Recombination rate variation shapes barriers to introgression across butterfly genomes. PLOS Biology **17**(2), 2006288. doi:10.1371/journal.pbio.2006288
17. Haenel, Q., Laurentino, T.G., Roesti, M., Berner, D.: Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. Molecular Ecology **27**(11), 2477–2497. doi:10.1111/mec.14699
18. Talla, V., Soler, L., Kawakami, T., Dincă, V., Vila, R., Friberg, M., Wiklund, C., Backström, N.: Dissecting the Effects of Selection and Mutation on Genetic Diversity in Three Wood White (Leptidea) Butterfly Species. Genome Biology and Evolution **11**(10), 2875–2886. doi:10.1093/gbe/evz212
19. Talavera, G., Bataille, C., Benyamini, D., Gascoigne-Pees, M., Vila, R.: Round-trip across the Sahara: Afrotropical Painted Lady butterflies recolonize the Mediterranean in early spring. Biology Letters **14**(6), 20180274. doi:10.1098/rsbl.2018.0274
20. Chapman, J.W., Reynolds, D.R., Wilson, K.: Long-range seasonal migration in insects: Mechanisms, evolutionary drivers and ecological consequences. Ecology Letters **18**(3), 287–302. doi:10.1111/ele.12407
21. Talavera, G., Vila, R.: Discovery of mass migration and breeding of the painted lady butterfly Vanessa cardui in the Sub-Sahara: The Europe–Africa migration revisited. Biological Journal of the Linnean Society **2016**(n/a). doi:10.1111/bij.12873
22. Menchetti, M., Guéguen, M., Talavera, G.: Spatio-temporal ecological niche modelling of multigenerational insect migrations. Proceedings of the Royal Society B: Biological Sciences **286**(1910), 20191583 (2019). doi:10.1098/rspb.2019.1583
23. Garcìa-Berro, A., Talla, V., Vila, R., Wai, H.K., Shipilina, D., Chan, K.G., Pierce, N.E., Backström, N., Talavera, G.: Genomic demographic inference shows migratory butterflies display higher heterozygosity and long-term effective population size. (in prep)
24. Celorio-Mancera, M.d.l.P., Wheat, C.W., Huss, M., Vezzi, F., Neethiraj, R., Reimegård, J., Nylin, S., Janz, N.: Evolutionary history of host use, rather than plant phylogeny, determines gene expression in a generalist butterfly. BMC Evolutionary Biology **16**(1), 59. doi:10.1186/s12862-016-0627-y
25. Ackery, P.R.: Hostplants and classification: a review of nymphalid butterflies. Biological Journal of the Linnean Society **33**(2), 95–203 (1988). doi:10.1111/j.1095-8312.1988.tb00446.x
26. Merlin, C., Liedvogel, M.: The genetics and epigenetics of animal migration and orientation: Birds, butterflies and beyond. JOURNAL OF EXPERIMENTAL BIOLOGY **222**. doi:10.1242/jeb.191890
27. Lohse, K., Wright, C., Talavera, G., García-Berro, A., Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium: The genome sequence of the painted lady, Vanessa cardui Linnaeus 1758. Wellcome Open Research **6**, 324 (2021). doi:10.12688/wellcomeopenres.17358.1
28. Connahs, H., Rhen, T., Simmons, R.B.: Transcriptome analysis of the painted lady butterfly, Vanessa cardui during wing color pattern development. BMC Genomics **17**(1), 270. doi:10.1186/s12864-016-2586-5
29. Zhang, L., Steward, R.A., Wheat, C.W., Reed, R.D.: High-Quality Genome Assembly and Comprehensive Transcriptome of the Painted Lady Butterfly Vanessa cardui. Genome Biology and Evolution **13**(7). doi:10.1093/gbe/evab145
30. Davey, J.W., Barker, S.L., Rastas, P.M., Pinharanda, A., Martin, S.H., Durbin, R., McMillan, W.O., Merrill, R.M., Jiggins, C.D.: No evidence for maintenance of a sympatric Heliconius species barrier by chromosomal inversions. Evolution Letters **1**(3), 138–154. doi:10.1002/evl3.12
31. Ahola, V., Lehtonen, R., Somervuo, P., Salmela, L., Koskinen, P., Rastas, P., Välimäki, N., Paulin, L., Kvist, J., Wahlberg, N., Tanskanen, J., Hornett, E.A., Ferguson, L.C., Luo, S., Cao, Z., de Jong, M.A., Duplouy, A., Smolander, O.-P., Vogel, H., McCoy, R.C., Qian, K., Chong, W.S., Zhang, Q., Ahmad, F., Haukka, J.K., Joshi, A., Salojärvi, J., Wheat, C.W., Grosse-Wilde, E., Hughes, D., Katainen, R., Pitkänen, E., Ylinen, J., Waterhouse, R.M., Turunen, M., Vähärautio, A., Ojanen, S.P., Schulman, A.H., Taipale, M., Lawson, D., Ukkonen, E., Mäkinen, V., Goldsmith, M.R., Holm, L., Auvinen, P., Frilander, M.J., Hanski, I.: The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. Nature Communications **5**(1), 4737. doi:10.1038/ncomms5737
32. Celorio-Mancera, M.d.l.P., Rastas, P., Steward, R.A., Nylin, S., Wheat, C.W.: Chromosome level assembly of the comma butterfly (Polygonia c-album). Genome Biology and Evolution. doi:10.1093/gbe/evab054
33. Gu, L., Reilly, P.F., Lewis, J.J., Reed, R.D., Andolfatto, P., Walters, J.R.: Dichotomy of Dosage Compensation along the Neo Z Chromosome of the Monarch Butterfly. Current Biology **29**(23), 4071–40773. doi:10.1016/j.cub.2019.09.056. 31735674
34. Li, X., Fan, D., Zhang, W., Liu, G., Zhang, L., Zhao, L., Fang, X., Chen, L., Dong, Y., Chen, Y., Ding, Y., Zhao, R., Feng, M., Zhu, Y., Feng, Y., Jiang, X., Zhu, D., Xiang, H., Feng, X., Li, S., Wang, J., Zhang, G., Kronforst, M.R., Wang, W.: Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies. Nature Communications **6**(1), 8212. doi:10.1038/ncomms9212
35. Smolander, O.-P., Blande, D., Ahola, V., Rastas, P., Tanskanen, J., Kammonen, J.I., Oostra, V., Pellegrini, L., Ikonen, S., Dallas, T., DiLeo, M.F., Duplouy, A., Duru, I.C., Halimaa, P., Kahilainen, A., Kuwar, S.S., Kärenlampi, S.O., Lafuente, E., Luo, S., Makkonen, J., Nair, A., de la Paz Celorio-Mancera, M., Pennanen, V., Ruokolainen, A., Sundell, T., Tervahauta, A.I., Twort, V., van Bergen, E., Österman-Udd, J., Paulin, L., Frilander, M.J., Auvinen, P., Saastamoinen, M.: 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**(1), 097. doi:10.1093/gigascience/giab097. 35022701
36. Yang, J., Wan, W., Xie, M., Mao, J., Dong, Z., Lu, S., He, J., Xie, F., Liu, G., Dai, X., Chang, Z., Zhao, R., Zhang, R., Wang, S.,

Zhang, Y., Zhang, W., Wang, W., Li, X.: Chromosome-level reference genome assembly and gene editing of the dead-leaf butterfly *Kallima inachus*. Molecular Ecology Resources **20**(4), 1080–1092. doi:10.1111/1755-0998.13185
37. Hill, J., Rastas, P., Hornett, E.A., Neethiraj, R., Clark, N., Morehouse, N., de la Paz Celorio-Mancera, M., Cols, J.C., Dircksen, H., Meslin, C., Keehnen, N., Pruisscher, P., Sikkink, K., Vives, M., Vogel, H., Wiklund, C., Woronik, A., Boggs, C.L., Nylin, S., Wheat, C.W.: Unprecedented reorganization of holocentric chromosomes provides insights into the enigma of lepidopteran chromosome evolution. Science Advances **5**(6), 3648. doi:10.1126/sciadv.aau3648
38. Van Belleghem, S.M., Rastas, P., Papanicolaou, A., Martin, S.H., Arias, C.F., Supple, M.A., Hanly, J.J., Mallet, J., Lewis, J.J., Hines, H.M., Ruiz, M., Salazar, C., Linares, M., Moreira, G.R.P., Jiggins, C.D., Counterman, B.A., McMillan, W.O., Papa, R.: Complex modular architecture around a simple toolkit of wing pattern genes. Nature Ecology & Evolution **1**(3), 1–12. doi:10.1038/s41559-016-0052
39. Lukhtanov, V.: The blue butterfly Polyommatus (Plebicula) atlanticus (Lepidoptera, Lycaenidae) holds the record of the highest number of chromosomes in the non-polyploid eukaryotic organisms. Comparative Cytogenetics **9**(4), 683–690. doi:10.3897/CompCytogen.v9i4.5760
40. de Vos, J.M., Augustijnen, H., Bätscher, L., Lucek, K.: Speciation through chromosomal fusion and fission in Lepidoptera. Philosophical Transactions of the Royal Society B: Biological Sciences **375**(1806), 20190539. doi:10.1098/rstb.2019.0539
41. Challi, R.J., Kumar, S., Dasmahapatra, K.K., Jiggins, C.D., Blaxter, M.: Lepbase: The Lepidopteran genome database, 056994. doi:10.1101/056994
42. Li, F., Zhao, X., Li, M., He, K., Huang, C., Zhou, Y., Li, Z., Walters, J.R.: Insect genomes: Progress and challenges. Insect Molecular Biology **28**(6), 739–758. doi:10.1111/imb.12599
43. Mita, K., Kasahara, M., Sasaki, S., Nagayasu, Y., Yamada, T., Kanamori, H., Namiki, N., Kitagawa, M., Yamashita, H., Yasukochi, Y., Kadono-Okuda, K., Yamamoto, K., Ajimura, M., Ravikumar, G., Shimomura, M., Nagamura, Y., Shin-i, T., Abe, H., Shimada, T., Morishita, S., Sasaki, T.: The Genome Sequence of Silkworm, Bombyx mori. DNA Research **11**(1), 27–35. doi:10.1093/dnares/11.1.27
44. Bachtrog, D.: Y chromosome evolution: Emerging insights into processes of Y chromosome degeneration. Nature reviews. Genetics **14**(2), 113–124. doi:10.1038/nrg3366. 23329112
45. Lewis, J.J., Cicconardi, F., Martin, S.H., Reed, R.D., Danko, C.G., Montgomery, S.H.: The Dryas iulia Genome Supports Multiple Gains of a W Chromosome from a B Chromosome in Butterflies. Genome Biology and Evolution **13**(7). doi:10.1093/gbe/evab128
46. Fraïsse, C., Picard, M.A.L., Vicoso, B.: The deep conservation of the Lepidoptera Z chromosome suggests a non-canonical origin of the W. Nature Communications **8**(1), 1486. doi:10.1038/s41467-017-01663-5
47. Lipinski, K.J., Farslow, J.C., Fitzpatrick, K.A., Lynch, M., Katju, V., Bergthorsson, U.: High Spontaneous Rate of Gene Duplication in Caenorhabditis elegans. Current biology : CB **21**(4), 306–310. doi:10.1016/j.cub.2011.01.026. 21295484
48. Loehlin, D.W., Carroll, S.B.: Expression of tandem gene duplicates is often greater than twofold. Proceedings of the National Academy of Sciences **113**(21), 5988–5992. doi:10.1073/pnas.1605886113. 27162370
49. Emerson, J.J., Cardoso-Moreira, M., Borevitz, J.O., Long, M.: Natural Selection Shapes Genome-Wide Patterns of Copy-Number Polymorphism in Drosophila melanogaster. Science **320**(5883), 1629–1631. doi:10.1126/science.1158078
50. Shields, O.: Journal of the Lepidopterists' Society. `https://www.biodiversitylibrary.org/item/128069`
51. HOSTS, h.n.a.u.-s.d.: HOSTS - a Database of the World's Lepidopteran Hostplants. `https://www.nhm.ac.uk/our-science/data/hostplants/search/index.dsml`
52. Nylin, S., Slove, J., Janz, N.: HOST PLANT UTILIZATION, HOST RANGE OSCILLATIONS AND DIVERSIFICATION IN NYMPHALID BUTTERFLIES: A PHYLOGENETIC INVESTIGATION: HOST RANGE OSCILLATIONS IN BUTTERFLIES. Evolution **68**(1), 105–124. doi:10.1111/evo.12227
53. Kang, L., Chen, X., Zhou, Y., Liu, B., Zheng, W., Li, R., Wang, J., Yu, J.: The analysis of large-scale gene expression correlated to the phase changes of the migratory locust. Proceedings of the National Academy of Sciences **101**(51), 17611–17615. doi:10.1073/pnas.0407753101
54. Zhu, H., Gegear, R.J., Casselman, A., Kanginakudru, S., Reppert, S.M.: Defining behavioral and molecular differences between summer and migratory monarch butterflies. BMC Biology **7**(1), 14. doi:10.1186/1741-7007-7-14
55. Landys, M.M., Piersma, T., Guglielmo, C.G., Jukema, J., Ramenofsky, M., Wingfield, J.C.: Metabolic profile of long-distance migratory flight and stopover in a shorebird. Proceedings of the Royal Society B: Biological Sciences **272**(1560), 295–302. doi:10.1098/rspb.2004.2952. 15705555
56. Murata, M., Tojo, S.: Utilization of lipid for flight and reproduction in Spodoptera litura (Lepidoptera: Noctuidae). EJE **99**(2), 221–224. doi:10.14411/eje.2002.031
57. Srygley, R.B., Dudley, R.: Optimal strategies for insects migrating in the flight boundary layer: Mechanisms and consequences. Integrative and Comparative Biology **48**(1), 119–133. doi:10.1093/icb/icn011
58. Weber, J.-M.: The physiology of long-distance migration: Extending the limits of endurance metabolism. Journal of Experimental Biology **212**(5), 593–597. doi:10.1242/jeb.015024
59. Breeschoten, T., van der Linden, C.F.H., Ros, V.I.D., Schranz, M.E., Simon, S.: Expanding the Menu: Are Polyphagy and Gene Family Expansions Linked across Lepidoptera? Genome Biology and Evolution **14**(1), 283. doi:10.1093/gbe/evab283
60. Hatfield, M.J., Umans, R.A., Hyatt, J.L., Edwards, C.C., Wierdl, M., Tsurkan, L., Taylor, M.R., Potter, P.M.: Carboxylesterases: General detoxifying enzymes. Chemico-biological interactions **259**, 327–331. doi:10.1016/j.cbi.2016.02.011. 26892220
61. Nagare, M., Ayachit, M., Agnihotri, A., Schwab, W., Joshi, R.: Glycosyltransferases: The multifaceted enzymatic regulator in insects. Insect Molecular Biology **30**(2), 123–137. doi:10.1111/imb.12686. 33263941
62. Huang, F.-F., Chai, C.-L., Zhang, Z., Liu, Z.-H., Dai, F.-Y., Lu, C., Xiang, Z.-H.: The UDP-glucosyltransferase multigene family in Bombyx mori. BMC Genomics **9**, 563. doi:10.1186/1471-2164-9-563. 19038024
63. Luque, T., Okano, K., O'Reilly, D.R.: Characterization of a novel silkworm (Bombyx mori) phenol UDP-glucosyltransferase. European Journal of Biochemistry **269**(3), 819–825. doi:10.1046/j.0014-2956.2001.02723.x
64. Hara, S., Yamakawa, M.: Moricin, a novel type of antibacterial peptide isolated from the silkworm, Bombyx mori. The Journal of Biological Chemistry **270**(50), 29923–29927. doi:10.1074/jbc.270.50.29923. 8530391
65. Zhan, S., Zhang, W., Niitepõld, K., Hsu, J., Haeger, J.F., Zalucki, M.P., Altizer, S., de Roode, J.C., Reppert, S.M., Kronforst, M.R.: The genetics of monarch butterfly migration and warning colouration. Nature **514**(7522), 317–321. doi:10.1038/nature13812
66. Chowdhury, S., Fuller, R.A., Dingle, H., Chapman, J.W., Zalucki, M.P.: Migration in butterflies: A global overview. Biological Reviews **96**(4), 1462–1483. doi:10.1111/brv.12714
67. Wahlberg, N., Rubinoff, D.: Vagility across Vanessa (Lepidoptera: Nymphalidae): Mobility in butterfly species does not inhibit the formation and persistence of isolated sister taxa. Systematic Entomology **36**(2), 362–370. doi:10.1111/j.1365-3113.2010.00566.x
68. Wieczorek, H., Beyenbach, K.W., Huss, M., Vitavska, O.: Vacuolar-type proton pumps in insect epithelia. The Journal of Experimental Biology **212**(11), 1611–1619. doi:10.1242/jeb.030007. 19448071
69. Beldade, P., Saenko, S.V., Pul, N., Long, A.D.: A Gene-Based Linkage Map for Bicyclus anynana Butterflies Allows for a Comprehensive Analysis of Synteny with the Lepidopteran Reference Genome. PLoS Genetics **5**(2), 1000366. doi:10.1371/journal.pgen.1000366
70. Rosser, N., Edelman, N.B., Queste, L.M., Nelson, M., Seixas, F., Dasmahapatra, K.K., Mallet, J.: Complex basis of hybrid female sterility and Haldane's rule in Heliconius butterflies: Z-linkage and

epistasis. Molecular Ecology (31). doi:10.1111/mec.16272
71. Tunström, K., Woronik, A., Hanly, J.J., Rastas, P., Chichvarkhin, A., Warren, A.D., Kawahara, A., Schoville, S.D., Ficarrotta, V., Porter, A.H., Watt, W.B., Martin, A., Wheat, C.W.: A complex interplay between balancing selection and introgression maintains a genus-wide alternative life history strategy, 2021–0520445023. doi:10.1101/2021.05.20.445023
72. Yamamoto, K., Nohata, J., Kadono-Okuda, K., Narukawa, J., Sasanuma, M., Sasanuma, S.-i., Minami, H., Shimomura, M., Suetsugu, Y., Banno, Y., Osoegawa, K., de Jong, P.J., Goldsmith, M.R., Mita, K.: A BAC-based integrated linkage map of the silkworm Bombyx mori. Genome Biology **9**(1), 21. doi:10.1186/gb-2008-9-1-r21
73. Yasukochi, Y.: A Dense Genetic Map of the Silkworm, Bombyx mori, Covering All Chromosomes Based on 1018 Molecular Markers. Genetics **150**(4), 1513–1525. doi:10.1093/genetics/150.4.1513
74. Jiggins, C.D., Mavarez, J., Beltrán, M., McMillan, W.O., Johnston, J.S., Bermingham, E.: A Genetic Linkage Map of the Mimetic Butterfly Heliconius melpomene. Genetics **171**(2), 557–570. doi:10.1534/genetics.104.034686. 15489522
75. Tobler, A., Kapan, D., Flanagan, N.S., Gonzalez, C., Peterson, E., Jiggins, C.D., Johntson, J.S., Heckel, D.G., McMillan, W.O.: First-generation linkage map of the warningly colored butterfly Heliconius erato. Heredity **94**(4), 408–417. doi:10.1038/sj.hdy.6800619
76. Kawakami, T., Mugal, C.F., Suh, A., Nater, A., Burri, R., Smeds, L., Ellegren, H.: 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**(16), 4158–4172. doi:10.1111/mec.14197
77. Scherthan, H., Weich, S., Schwegler, H., Heyting, C., Härle, M., Cremer, T.: 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**(5), 1109–1125. doi:10.1083/jcb.134.5.1109
78. Smith, G.R., Nambiar, M.: New Solutions to Old Problems: Molecular Mechanisms of Meiotic Crossover Control. Trends in genetics : TIG **36**(5), 337–346. doi:10.1016/j.tig.2020.02.002. 32294414
79. Kent, T.V., Uzunović, J., Wright, S.I.: Coevolution between transposable elements and recombination. Philosophical Transactions of the Royal Society B: Biological Sciences **372**(1736), 20160458. doi:10.1098/rstb.2016.0458
80. Rizzon, C., Marais, G., Gouy, M., Biémont, C.: Recombination Rate and the Distribution of Transposable Elements in the Drosophila melanogaster Genome. Genome Research **12**(3), 400–407. doi:10.1101/gr.210802. 11875027
81. Deininger, P.L., Batzer, M.A.: Alu Repeats and Human Disease. Molecular Genetics and Metabolism **67**(3), 183–193. doi:10.1006/mgme.1999.2864
82. Apuli, R.-P., Bernhardsson, C., Schiffthaler, B., Robinson, K.M., Jansson, S., Street, N.R., Ingvarsson, P.K.: Inferring the Genomic Landscape of Recombination Rate Variation in European Aspen (*Populus tremula*). G3: Genes—Genomes—Genetics **10**(1), 299–309. doi:10.1534/g3.119.400504
83. Kawakami, T., Smeds, L., Backström, N., Husby, A., Qvarnström, A., Mugal, C.F., Olason, P., Ellegren, H.: A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. Molecular Ecology **23**(16), 4035–4058. doi:10.1111/mec.12810
84. Boman, J., Mugal, C.F., Backström, N.: The Effects of GC-Biased Gene Conversion on Patterns of Genetic Diversity among and across Butterfly Genomes. Genome Biology and Evolution **13**(5), 064. doi:10.1093/gbe/evab064
85. Aljanabi, S.: Universal and rapid salt-extraction of high quality genomic DNA for PCR- based techniques. Nucleic Acids Research **25**(22), 4692–4693. doi:10.1093/nar/25.22.4692
86. Bioinformatics, B., Andrew, S.: FastQC A Quality Control Tool for High Throughput Sequence Data. `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`
87. Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., Cresko, W.A.: Stacks: An analysis tool set for population genomics. Molecular Ecology **22**(11), 3124–3140. doi:10.1111/mec.12354
88. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. doi:10.48550/arXiv.1303.3997
89. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup: The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) **25**(16), 2078–2079. doi:10.1093/bioinformatics/btp352. 19505943
90. Okonechnikov, K., Conesa, A., García-Alcalde, F.: Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics, 566. doi:10.1093/bioinformatics/btv566
91. Rastas, P.: Lep-MAP3: Robust linkage mapping even for low-coverage whole genome sequencing data. Bioinformatics **33**(23), 3726–3732. doi:10.1093/bioinformatics/btx494
92. Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G.: QUAST: Quality assessment tool for genome assemblies. Bioinformatics **29**(8), 1072–1075. doi:10.1093/bioinformatics/btt086
93. Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.-h., Jin, H., Marler, B., Guo, H., Kissinger, J.C., Paterson, A.H.: MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Research **40**(7), 49. doi:10.1093/nar/gkr1293. 22217600
94. Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A.: Circos: An information aesthetic for comparative genomics. Genome Research **19**(9), 1639–1645. doi:10.1101/gr.092759.109
95. Holt, C., Yandell, M.: MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics **12**(1), 491. doi:10.1186/1471-2105-12-491
96. Smit, A., Hubley, R., Green, P.: RepeatMasker Open-4.0. `<http://www.repeatmasker.org>`
97. Korf, I.: Gene finding in novel genomes. BMC Bioinformatics **5**(1), 59. doi:10.1186/1471-2105-5-59
98. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., Hunter, S.: InterProScan 5: Genome-scale protein function classification. Bioinformatics **30**(9), 1236–1240. doi:10.1093/bioinformatics/btu031
99. Bao, W., Kojima, K.K., Kohany, O.: Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA **6**(1), 11. doi:10.1186/s13100-015-0041-9
100. Zhan, S., Reppert, S.M.: MonarchBase: The monarch butterfly genome database. Nucleic Acids Research **41**(D1), 758–763. doi:10.1093/nar/gks1057
101. Quinlan, A.R., Hall, I.M.: BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6), 841–842. doi:10.1093/bioinformatics/btq033
102. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., von Mering, C., Bork, P.: eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research **47**(D1), 309–314. doi:10.1093/nar/gky1085
103. Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., Zdobnov, E.M.: BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution **38**(10), 4647–4654. doi:10.1093/molbev/msab199
104. Emms, D.M., Kelly, S.: OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biology **20**(1), 238. doi:10.1186/s13059-019-1832-y
105. Librado, P., Vieira, F.G., Rozas, J.: BadiRate: Estimating family turnover rates by likelihood-based methods. Bioinformatics **28**(2), 279–281. doi:10.1093/bioinformatics/btr623
106. Huerta-Cepas, J., Serra, F., Bork, P.: ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Molecular Biology

and Evolution **33**(6), 1635–1638. doi:10.1093/molbev/msw046
107. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6), 716–723. doi:10.1109/TAC.1974.1100705
108. Wolfe, D., Dudek, S., Ritchie, M.D., Pendergrass, S.A.: Visualizing genomic information across chromosomes with PhenoGram. BioData Mining **6**(1), 18. doi:10.1186/1756-0381-6-18
109. Alexa, A., Rahnenfuhrer, J.: topGO: enrichment Analysis for Gene Ontology. doi:10.18129/B9.bioc.topGO. Bioconductor version: Release (3.13). `https://bioconductor.org/packages/topGO/`
110. (2021)., R.C.T.: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.r-project.org/index.html`
111. Alexa, A., Rahnenführer, J., Lengauer, T.: Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics **22**(13), 1600–1607. doi:10.1093/bioinformatics/btl140
112. Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) **57**(1), 289–300. 2346101
113. Rezvoy, C., Charif, D., Gueguen, L., Marais, G.A.B.: MareyMap: An R-based tool with graphical interface for estimating recombination rates. Bioinformatics **23**(16), 2188–2189. doi:10.1093/bioinformatics/btm315
114. Bauer, D.F.: Constructing Confidence Sets Using Rank Statistics. Journal of the American Statistical Association **67**(339), 687–690. doi:10.1080/01621459.1972.10481279
115. Best, D.J., Roberts, D.E.: Algorithm AS 89: The Upper Tail Probabilities of Spearman's Rho. Applied Statistics **24**(3), 377. doi:10.2307/2347111. 2347111
116. Royston, J.P.: An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. Applied Statistics **31**(2), 115. doi:10.2307/2347973. 10.2307/2347973
117. Wickham, H.: Ggplot2: Elegant Graphics for Data Analysis. Use R! Springer-Verlag. doi:10.1007/978-0-387-98141-3. `https://www.springer.com/gp/book/9780387981413`