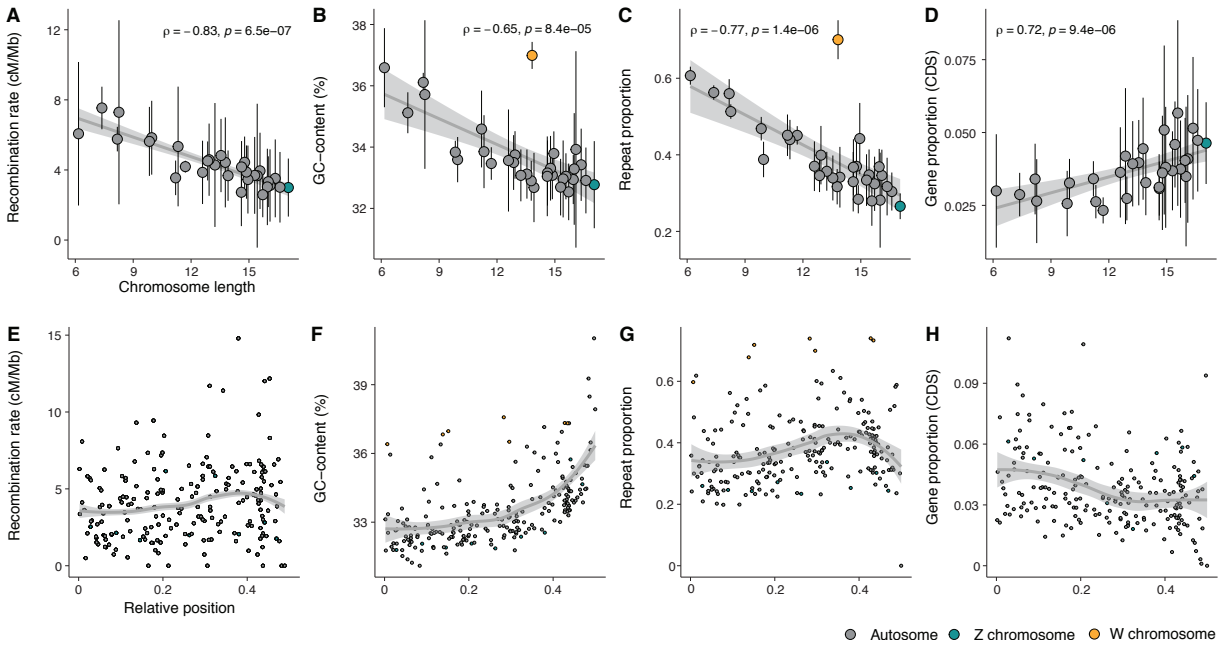

Figure 4 Top row: Associations between chromosome length and A) recombination rate, B) base composition, C) repeat and D) gene proportions. Chromosome length is given in megabases (Mb). Bottom row: Regional distribution of the recombination rate (E), base composition (F), repeat (G) and gene (H) density 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.

the positive association with the proportion of SINEs (Estimate 1.59, p-value = 9.75×10^{-04}) and the negative association with chromosome size (Estimate -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

General

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 (Celorio-Mancera et al., 2021; Gu et al., 2019; Li et al., 2015; Smolander et al., 2022; Yang et al., 2020), 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 (Davey et al., 2017; Hill et al., 2019; Van Belleghem et al., 2017). 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 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

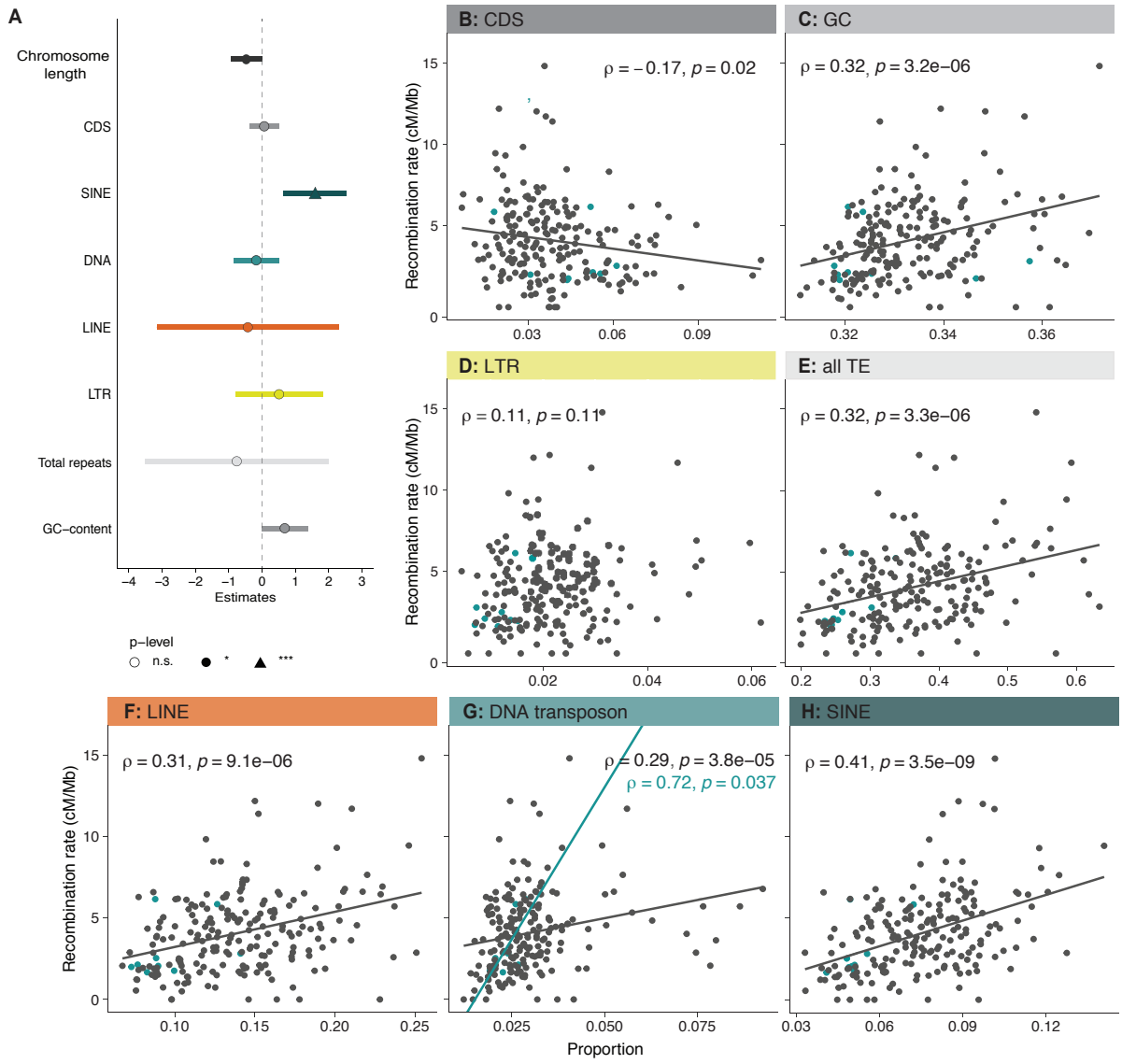

Figure 5 Correlation between recombination rate and density of genomic features. A) Summary of the linear model with regional recombination rate as the response variable. Each explanatory variable in the model is listed along the Y-axis and the relative estimated effect (X-axis), and error intervals 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 p-values for significant analyzes are given. Gray dots/lines indicate autosomal regions and turquoise dots/lines Z-chromosome linked regions

painted lady developed 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 (de Vos et al., 2020; Lukhtanov, 2015). In agreement with previous data (Zhang et al., 2021), 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 (Davey et al., 2017). Hence, similar to other nymphalid butterflies, the painted lady has retained

the inferred ancestral lepidopteran karyotype (Ahola et al., 2014). The annotation procedure revealed that the painted lady harbors a gene set (n = 13,161) close to the suggested core set in Lepidoptera (Challi et al., 2016; Li et al., 2019) and a relatively low overall TE content. However, the TE content was significantly higher and the gene density lower on smaller chromosomes.

A clear outlier for gene density and TE content was the W-chromosome. While having a size of equal to an average autosome, the W-chromosome demonstrated very specific features; 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 (Lewis et al., 2021; Mita et al., 2004). 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 on the painted lady W-chromosome was also observed in the silkmoth (Abe et al., 2008; Mita et al., 2004), and is likely a consequence of the degradation process of the non-recombining sex-chromosome (Bachtrog, 2013). 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 (Bachtrog, 2013). The Z-chromosome is generally highly conserved in Lepidoptera (Fraïsse et al., 2017) 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 (Lipinski et al., 2011). However, most gene duplication events are probably deleterious (Loehlin and Carroll, 2016) or effectively neutral, leading to a low probability of fixation of novel gene copies (Emerson et al., 2008). 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 (Garcìa-Berro et al., in prep), which translates to efficient selection against slightly deleterious variants. The majority of the 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 an extraordinary life-history and has become a quickly uprising complementary model organism for studying insect migration. Over most of the almost cosmopolitan distribution range (Shields, 1992), the painted ladies complete a multigenerational migratory circuit, where single individuals can migrate $> 4,000$ kilometers during lifetime (Talavera and Vila, n.d.). In contrast to other migratory butterflies like the monarch and the red admiral (*Vanessa atalanta*), the painted lady is non-diapausing (Shields, 1992). The genetic underpinnings of migratory behavior have only been preliminarily characterized for a handful of insect species (Kang et al., 2004; Zhu et al., 2009) 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 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 (Landys et al., 2005; Murata and Tojo, 2013; Srygley and Dudley, 2008; Weber, 2009), both the capacity to build up fat deposits and efficient sequestration of fatty acids have likely been under 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.

Long-range migrants 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 utilizes > 300 different larval host-plants in 11 plant families (Ackery, 1988; Celorio-Mancera et al., n.d.; Nylin et al., 2014). Two of the significantly expanded gene families in the painted lady (UDP-glycosyltransferase,

carboxylesterase) were associated with polyphagy and detoxification (Breeschoten et al., 2022; Hatfield et al., 2016; Nagare et al., 2021). The UDP-glycosyltransferase superfamily includes Lepidoptera-specific subfamilies associated with a variety of functions, such as affinity for plant secondary metabolites (Huang et al., 2008; Luque et al., 2002). In the painted lady larvae, one UDP-subfamily is upregulated in response to utilization of an extended range of hostplants (Celorio-Mancera et al., n.d.). 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 (Hara and Yamakawa, 1995), 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.

Previous investigation of the genetic basis of migratory behavior in the monarch butterfly identified candidate genes associated with orientation, chemoreception and regulation of the circadian clock (Zhan et al., 2014; Zhu et al., 2009). Migratory behavior has evolved independently multiple times within the Papilionoidea clade (Chowdhury et al., 2021) and in the *Vanessa* genus (Wahlberg and Rubinoff, 2011), 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. Significantly expanded gene families shared between the painted lady and the monarch were enriched for functions associated with various metabolic processes, defense against pathogens and neuronal activity, all of which can be associated with migratory behavior. One gene family with an especially pronounced expansion was vacuolar ATPases, ATP-dependent proton pumps, involved in membrane ion transport (Wieczorek et al., 2009). 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 of independent migratory and sedentary sister species, in combination with detailed population genetic analysis and functional verification 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 genetic drift and selection on levels of genetic diversity. Understanding how recombination breaks down linkage disequilibrium is also important for association studies aimed at coupling genetic variation to phenotypic traits. Despite their importance, detailed recombination maps are only available for a handful of butterfly species (Beldade et al., 2009; Celorio-Mancera et al., 2021; Davey et al., 2017; Rosser et al., 2022; Smolander et al., 2022; Tunström et al., 2021). In some butterfly species, linkage maps have been used to improve and/or verify the correctness of physical genome assemblies, but the recombination rate has not been assessed. 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 (Davey et al., 2017; Martin et al., 2019). This indicates that the painted lady genetic map accurately reflects the historical recombination landscape in the species.

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 (Yamamoto et al., 2008; Yasukochi, 1998) to 5.5 - 6.0 cM / Mb in different *Heliconius* species (Jiggins et al., 2005; Tobler et al., 2005). 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 (Haenel et al., 2018; Kawakami et al., 2017; Martin et al., 2019).

Butterflies and moths have holocentric chromosomes, i.e. they lack distinct centromere regions, which might lead to an expectation of a uniform distribution of recombination events. In the painted lady we observed a bimodal distribution of recombination events along chromosomes, with an increase of recombination rate away from the center and significant drop at the chromosome ends. This distribution is in agreement with previous observations, both in Lepidoptera and in other animals with different centromere types (Haenel et al., 2018; Martin et al., 2019). A possible explanation for this pattern is mechanical or tension interference between chiasmata when > 1 recombination event occurs on the same chromosome (Haenel et al., 2018). 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 (Barnes et al., 1995). An alternative explanation could be that synaptonemal complexes are directed towards the flanking regions, when the telomeres attach to the nuclear wall (Scherthan et al., 1996). The 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 (Smith and Nambiar, 2020).

Since recombination is directly associated with the efficacy of selection, we expect a negative correlation between the regional recombination rate and a number of repeats. Such associations have been observed in many organisms, although its degree varies to some extent (Kent et al., 2017; Rizzon et al., 2002). 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. SINE can mediate recombination, as has been described in humans (Deininger and Batzer, 1999), but we can not exclude other factors affecting both 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. One interesting observation was the radically different distribution of TE-classes on the W-chromosome in the painted lady, with a very low frequency of SINEs as compared to the autosomes and the Z-chromosome. The absence of SINEs on the non-recombining W-chromosome, and the strong positive association between SINE density and recombination rate on the autosomes and the Z-chromosome, hence suggests that SINEs might be able to hijack the recombination machinery and mediate their own proliferation via double-strand breaks.

In contrast with results from similar studies in other organism groups (Apuli et al., 2020; Kawakami et al., 2014), 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 (Boman et al., 2021). 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