Genomics

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

--Manuscript Draft--

Manuscript Number:	GEN-D-22-00413R1
Article Type:	Research Paper
Section/Category:	
Keywords:	genomics; recombination; linkage map; gene family; painted lady; Lepidoptera
Corresponding Author:	Daria Shipilina Uppsala University Uppsala, SWEDEN
First Author:	Daria Shipilina
Order of Authors:	Daria Shipilina
	Karin Näsvall
	Lars Höök
	Roger Vila
	Gerard Talavera
	Niclas Backström
Abstract:	Characterization of gene family expansions and crossing over is crucial for understanding how organisms adapt to the environment. Here, we develop a high-density linkage map and detailed genome annotation of the painted lady butterfly (Vanessa cardui) - a non-diapausing, highly polyphagous species famous for its long-distance migratory behavior and almost cosmopolitan distribution. Our results reveal a complex interplay between regional recombination rate variation, gene duplications and transposable element activity shaping the genome structure of the painted lady. We identify several lineage specific gene family expansions with functions mainly associated with protein and fat metabolism, detoxification, and defense against infection - critical processes for the painted lady's unique life-history. Furthermore, the detailed recombination maps allow us to characterize the regional recombination landscape, and reveal a strong effect of chromosome size on the recombination rate, a limited impact of GC-biased gene conversion and a positive association between recombination and short interspersed elements.
Suggested Reviewers:	Simon Martin simon.martin@ed.ac.uk
	Marjo Saastamonien marjo.saastamoinen@helsinki.fi
	Christine Merlin cmerlin@bio.tamu.edu
	Nick Grishin NICK.GRISHIN@UTSOUTHWESTERN.EDU
Opposed Reviewers:	
Response to Reviewers:	

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

UPPSALA
UNIVERSITET

Dear Editors,

Please find attached a revised version of our manuscript "Linkage mapping and genome annotation give novel insights into gene family expansions and regional recombination rate variation in the painted lady (*Vanessa cardui*) butterfly" attached to this submission.

We appreciate the prompt handling and the insightful comments from you and the two reviewers on the initial version of the paper.

Best regards,

Authors

Comments from editor Colum Walsh:
I have completed my evaluation of your manuscript. The reviewers recommend reconsideration of your manuscript following minor revision and modification. I invite you to resubmit your manuscript after addressing the comments below. Please resubmit your revised manuscript by Sep 01, 2022.

When revising your manuscript, please consider all issues mentioned in the reviewers' comments carefully: please outline every change made in response to their comments and provide suitable rebuttals for any comments not addressed. Please note that your revised submission may need to be re-reviewed.

Genomics values your contribution and I look forward to receiving your revised manuscript.

Author response: *Dear Editor Walsh. We appreciate the insightful comments from you and the reviewers and have done our best to address them all in the revised version. Please, find an updated version of the manuscript in the submission portal where all changes have been highlighted for straightforward re-evaluation.*

Reviewer comments:

Reviewer #1: The work by Shipilina et al on a recombination map for a butterfly species was a very well written and articulated the need and results of the research, and the results track previous work in Lepidoptera.

Author response: *We appreciate the overall positive notes on the manuscript and are grateful for the comments. The manuscript improved considerably after taking the comments into account.*

The manuscript would benefit from reworking part of the introduction so the ideas flow better for the generalist audience of Genomics: the second and third paragraphs can be shortened (even though the writing is very high quality) as the purpose of the manuscript is not to educate the reader on the evolutionary significance of gene duplications and TEs but to introduce the research results. Whilst this would replace a couple of the longer sentences in each paragraph, the text savings replaced with knowledge that will help the reader understand why the presented results are important such as value of discussing the TE and gene families. For example, my understanding from the manuscript's results was that the value of TEs here is mainly as markers rather than biological entities: undertaking an analysis of TE evolution across butterflies would be a completely separate endeavour but that's what I thought I was going to learn after reading the introduction!

Author response: *Thank you for the suggestion. The introduction has now been edited to better fit the scope of the journal and the audience. Specifically, the second and third paragraphs have been shortened as suggested and we now focus more on the relevance of the TE annotation and gene duplication analysis in general. Restructured sentences are highlighted in green.*

Further, explaining the holocentric nature of Lepidopteran chromosomes in the introduction would be vital for the reader to understand the constraints butterfly chromosomes face. Indeed, the first part of the results confirms this hypothesis and it would be valuable to the reader to know that these hypotheses have been voiced and are being explored (here and in previous work).

Author response: *We have now added a paragraph outlining the specifics of holocentricity and how that feature can affect genome evolution in general and recombination rate variation in particular. Section starting from "Such spatial variation…".*

I think the results are very robust but I wonder why GO terms were used to limit the subset (I am not a fan of the approach for non-models generally). One approach is to use sonic inParanoid to do a all vs all proteome comparison and identify gene family expansions regardless of GO terms. Given the results presented, if a grouping classification was necessary, I find the Enzyme classifications more robust as they are less prone to misclassications. However, I would not expect the authors to redo their analyses at this stage, merely to consider it and mention the issue of using GO as a nomenclature rather than a rough grouping.

Author response: *We agree with the reviewer, that using functional annotation for non-model organisms should be interpreted with caution. However, since the GO-term analysis was performed after orthogroup assignment with OrthoFinder, following evolutionary rate estimations with BadiRate, we think we avoided some of the potential biases. Additionally, we curated gene families using domain annotation (InterProScan) before evaluating enriched GO terms. We are also grateful for the suggestion to use the SonicParanoid suite and will consider it for potential future projects (Cosentino, Iwasaki, 2019).*

I'm glad to see the data to be made publicly available.

Minor comments

Placing - and discussing - both of the sex chromosomes in their own section before Gene Family Analysis would help those of us interested in the topic. I would appreciate a line in the discussion highlighting the difficulty of sequencing the W so that the reader is aware.

Author response: *The structure of the discussion has been changed accordingly. We have also added a short piece of text in the discussion where we address the problems with sequencing and assembling highly repetitive sex-limited chromosomes. It can be found in the beginning of the new "Sex-chromosome" section.*

I was not familiar with the argumentation in the sentence "The absence of SINEs on the ...". Is there any support that SINEs hijack replication machinery but either LTRs are more inefficient or the correlation is watered down by the other parts of their biology?

Author response: *Thank you for the feedback. We have now reformulated the section a bit to make our argumentation clearer. The hypothesis that SINEs may hijack the recombination process comes from our observation that the non-recombining W-chromosome has a very low density of SINEs. This is also supported by studies of Alu elements in humans (cited in main text; Batzer, M., and Deininger, P. 2002. "Alu repeats and human genomic diversity". Nat. Rev. Genet. 3, 370–379).*

Please consider

- "General" after discussion is not needed but other subheadings are needed as the info was quite packed.

Author response: *The discussion has now been broken up in smaller sections with sub-headers, as suggested.*

- changing "Additional curation" -> "manual curation"

Author response: *Changed accordingly.*

Reviewer #2: This is a fine piece of genome analysis work. The methodology is state of the art, and the results add considerably to lepidoptera genomics in general and to the painted lady in particular.

Initially I was a bit surprised that the authors based their work on a previously sequenced genome (Lohse et al., 2021), but that publication is basically a genome announcement. The current work goes well beyond a simple reannotation.

Author response: *Thank you. We are grateful for the very positive review of the manuscript.*

Revised Manuscript with Changes Marked

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

Characterization of gene family expansions and crossing over is crucial for understanding how organisms adapt to the environment. Here, we develop a high-density linkage map and detailed genome annotation of the painted lady butterfly (*Vanessa cardui*) - a non-diapausing, highly polyphagous species famous for its long-distance migratory behavior and almost cosmopolitan distribution. Our results reveal a complex interplay between regional recombination rate variation, gene duplications and transposable element activity shaping the genome structure of the painted lady. We identify several lineage specific gene family expansions. Their functions are mainly associated with protein and fat metabolism, detoxification, and defense against infection - critical processes for the painted lady's unique life-history. Furthermore, the detailed recombination maps allow us to characterize the regional recombination landscape, data that reveal a strong effect of chromosome size on the recombination rate, a limited impact of GC-biased gene conversion and a positive association between recombination and short interspersed elements.

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 (Henikoff, 1997; Ojeda-López et al., 2020; Zhang, 2003), and gene family expansions and contractions are important sources for generation of phenotypic diversity (Chen et al., 2013; Kondrashov, 2012). 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 (Schwander et al., 2014).

Since the spearheading work by McClintock (McClintock, 1956), transposable elements (TEs) have been acknowledged as major contributors to different types of evolutionary change in eukaryotes (Kazazian, 2004; Kidwell and Lisch, 1997). Transposable elements are capable of self-replication within the host genome and can mediate both small scale deletions and duplications, large scale chromosomal rearrangements (Kidwell and Lisch, 1997) and considerable genome size expansions (Podsiadlowski et al., 2021; Talla et al., 2017). In 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 (Peñalba and Wolf, 2020). Recombination also influences selection efficiency by directly preventing the accumulation of deleterious alleles (Müller's ratchet) and breaking the physical linkage

+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

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 (Stapley et al., 2017; Tiley and Burleigh, 2015). However, besides detailed recombination maps in the butterfly genus *Heliconius* (Martin et al., 2019), 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 (Haenel et al., 2018; Talla et al., 2019).

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 wide-spread of all butterfly species (Talavera et al., 2018), 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 (Chapman et al., 2015; Guerra et al., 2014). 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 (Menchetti et al., 2019; Talavera and Vila, n.d.). 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 (Garcìa-Berro et al., in prep). This could also be a consequence of the species' ability to utilize a wide range of host plants (Ackery, 1988; Celorio-Mancera et al., 2016). 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 (Merlin and Liedvogel, 2019).

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 (Lohse et al., 2021). 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 (Connahs et al., 2016; Zhang et al., 2021) 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 (Lohse et al., 2021) 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; $R = 0.91$-1.00, p-value $> 1.00 \times 10^{-04}$, 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).

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).

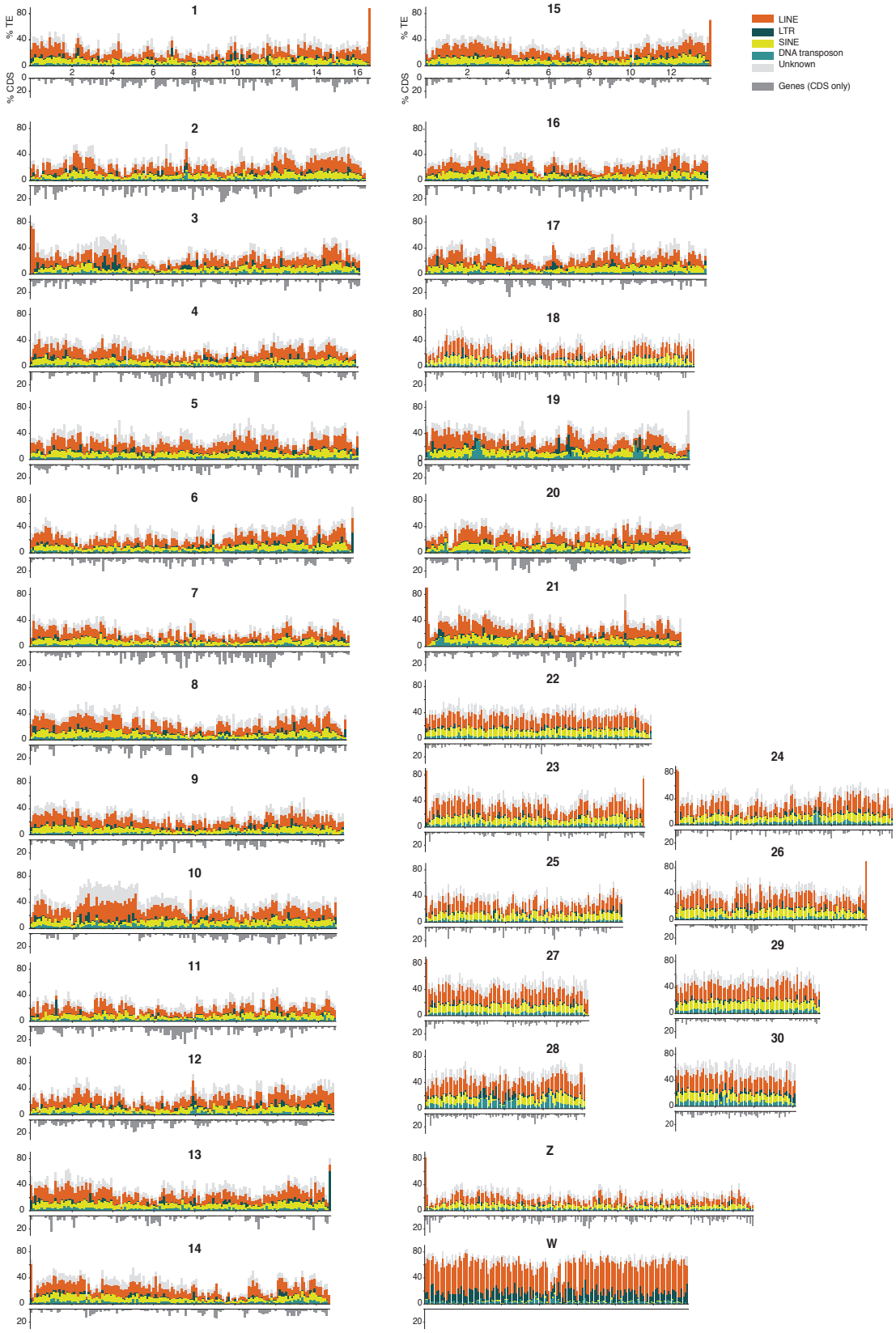

Table 1 Linkage map, genome assembly and annotation statistics

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 (Davey et al., 2017) could also be verified (Figure 2). In summary, this confirms that the painted lady karyotype is highly similar to the inferred ancestral butterfly karyotype (Ahola et al., 2014).

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).

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 (Spearman's rank correlation, $\rho = -0.83$, p-value $= 6.51 \times 10^{-07}$;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 (Figure 4A). Besides

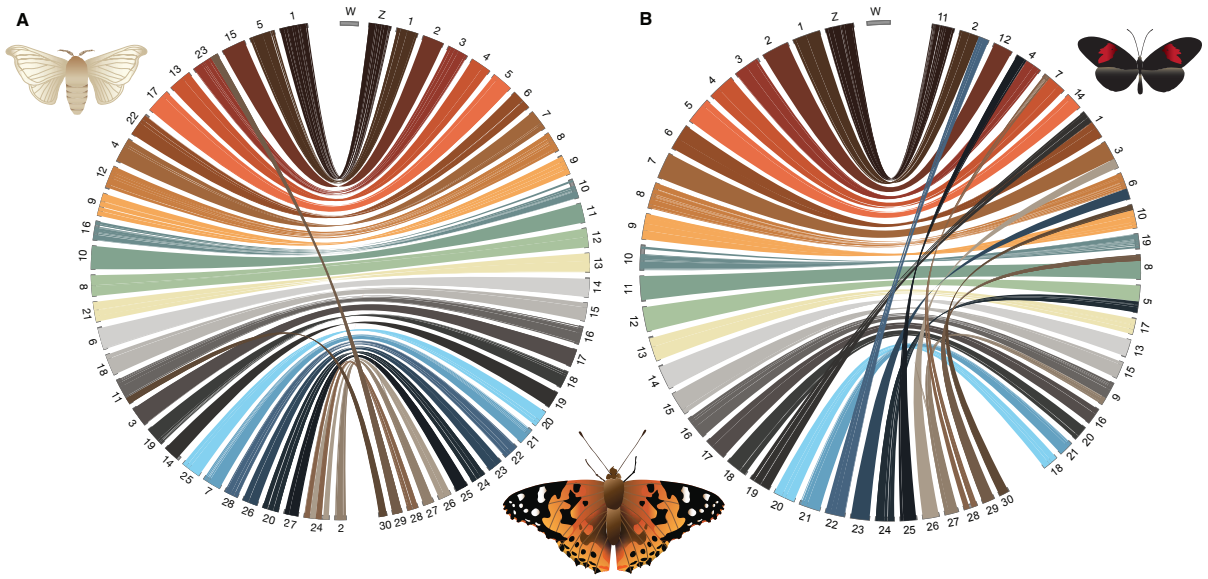

Figure 2 Synteny between the painted lady and a) the silkmoth (*Bombyx mori*) and b) the postman butterfly (*Heliconius melpomene*) chromosomes, respectively. The painted lady autosomes are sorted and named according to length and the sex chromosomes are indicated with Z and W. Orthologous genes are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

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 $= 8.35\times10^{-05}$) and repeat density ($\rho = 0.77$, p-value $= 1.37\times10^{-06}$), and a positive association with gene density ($\rho = 0.68$, p-value $= 2.63\times10^{-05}$)(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 ($\rho = 0.32$, p-value $= 3.68\times10^{-06}$). Gene density was homogeneous across chromosomes, with only a minor increase towards the chromosome center, and was negatively associated with the recombination rate ($\rho = $-0.19, p-value $= 7.27\times10^{-03}$). We found a significant positive association between the overall repeat proportion and the recombination rate ($\rho = 0.35$, p-value $= 3.48\times10^{-07}$; Figure 5), and this pattern was consistent for all repeat classes, but strongest for SINEs ($\rho = 0.42$, p-value $= 2.63\times10^{-10}$) and weakest for LTRs ($\rho = 0.14$, p-value $= 4.04\times10^{-02}$). The association between recombination rate and proportion of LTRs was, however, not significant when only including autosomes ($\rho = 0.11$, p-value $= 11.04\times10^{-02}$; 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

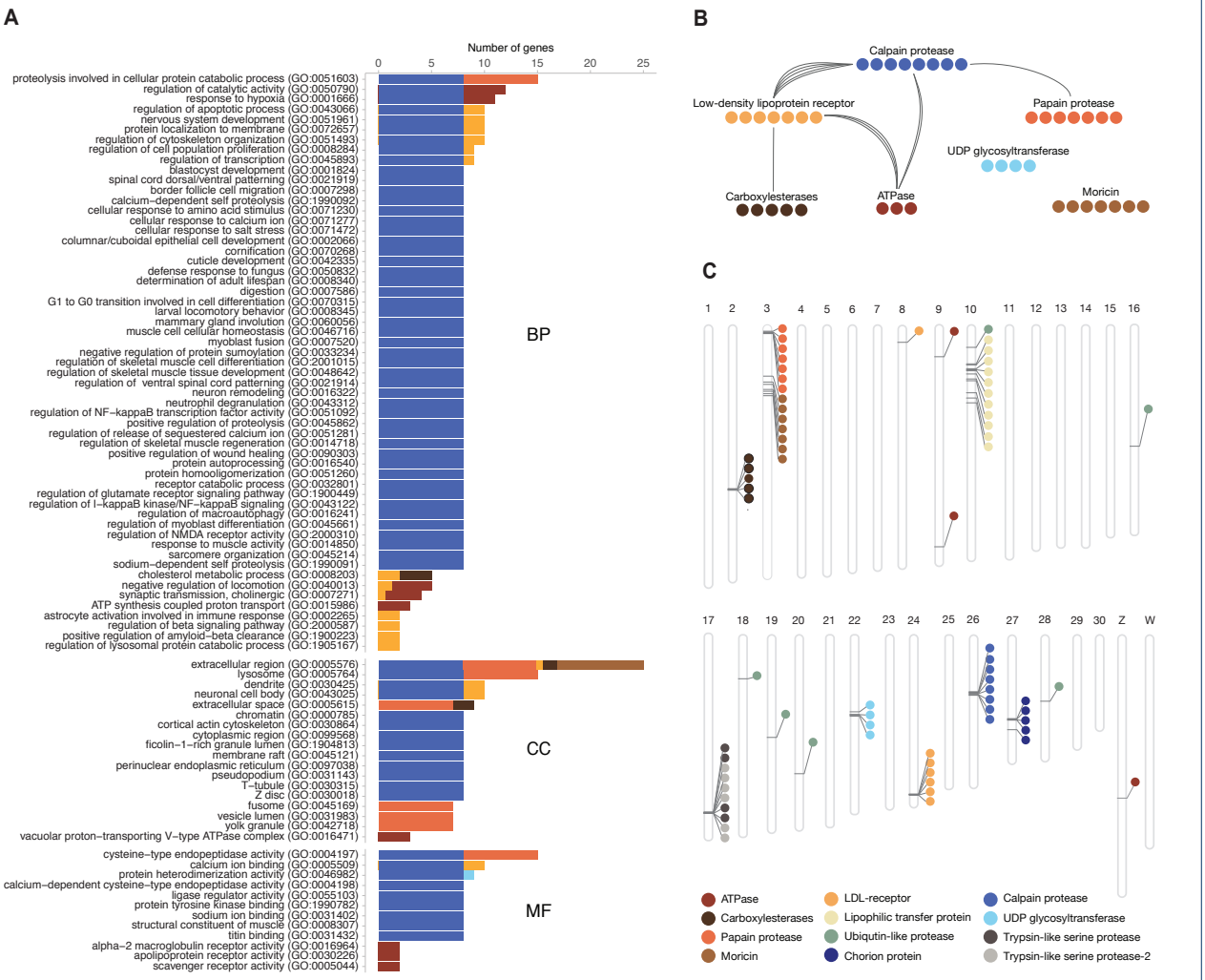

Figure 3 A) Significantly (p-value < 0.05 after FDR-correction) enriched gene ontology (GO) terms associated with expanded gene families in the painted lady. The bars show the number of genes associated with each GO-term. The different GO-categories are biological process (BP), cellular compartment (CC) 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.

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 $= 5.36\times10^{-09}$) 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 by the positive association with the proportion of SINEs (Estimate 1.59, p-value $= 9.75\times10^{-04}$) and the negative association with chromosome size (Estimate -0.48, p-value $= 4.88\times10^{-02}$; 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 $= 9.1\times10^{-01}$ - 8.0×10^{-01}; Figure S6).

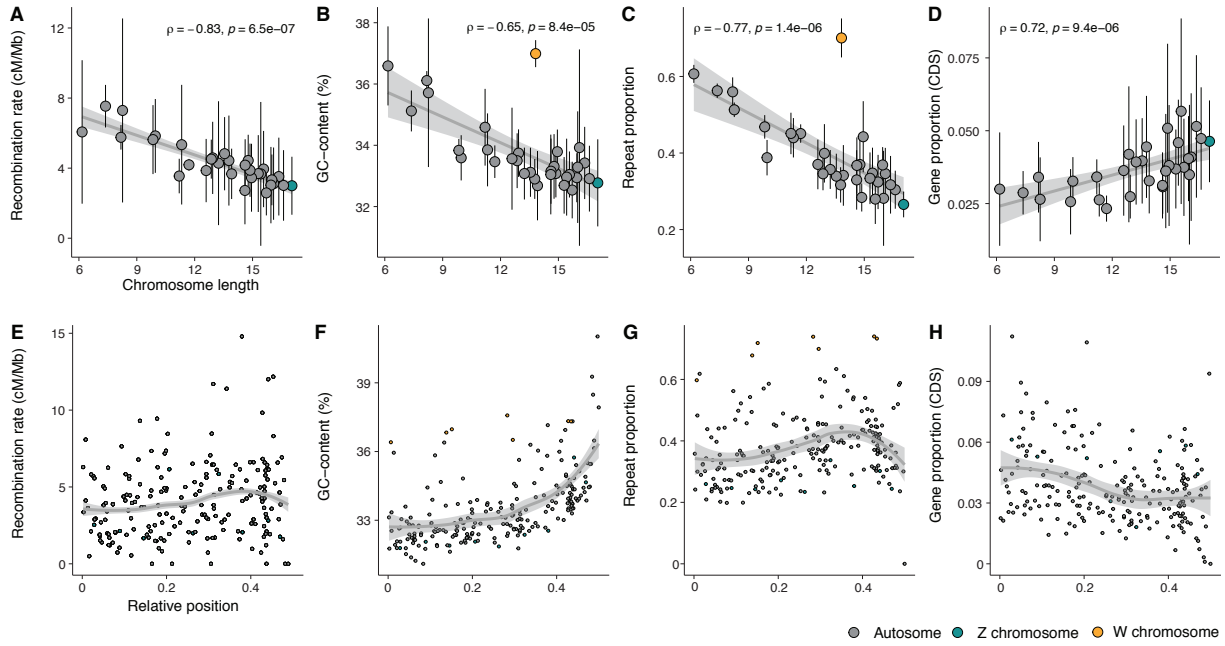

Figure 4 Top row: Associations between chromosome length and A) recombination rate, B) base composition, C) repeat and D) gene proportions. Chromosome length is given in megabases (Mb). Bottom row: Regional distribution of the recombination rate (E), base composition (F), repeat (G) and gene (H) density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

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 (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 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

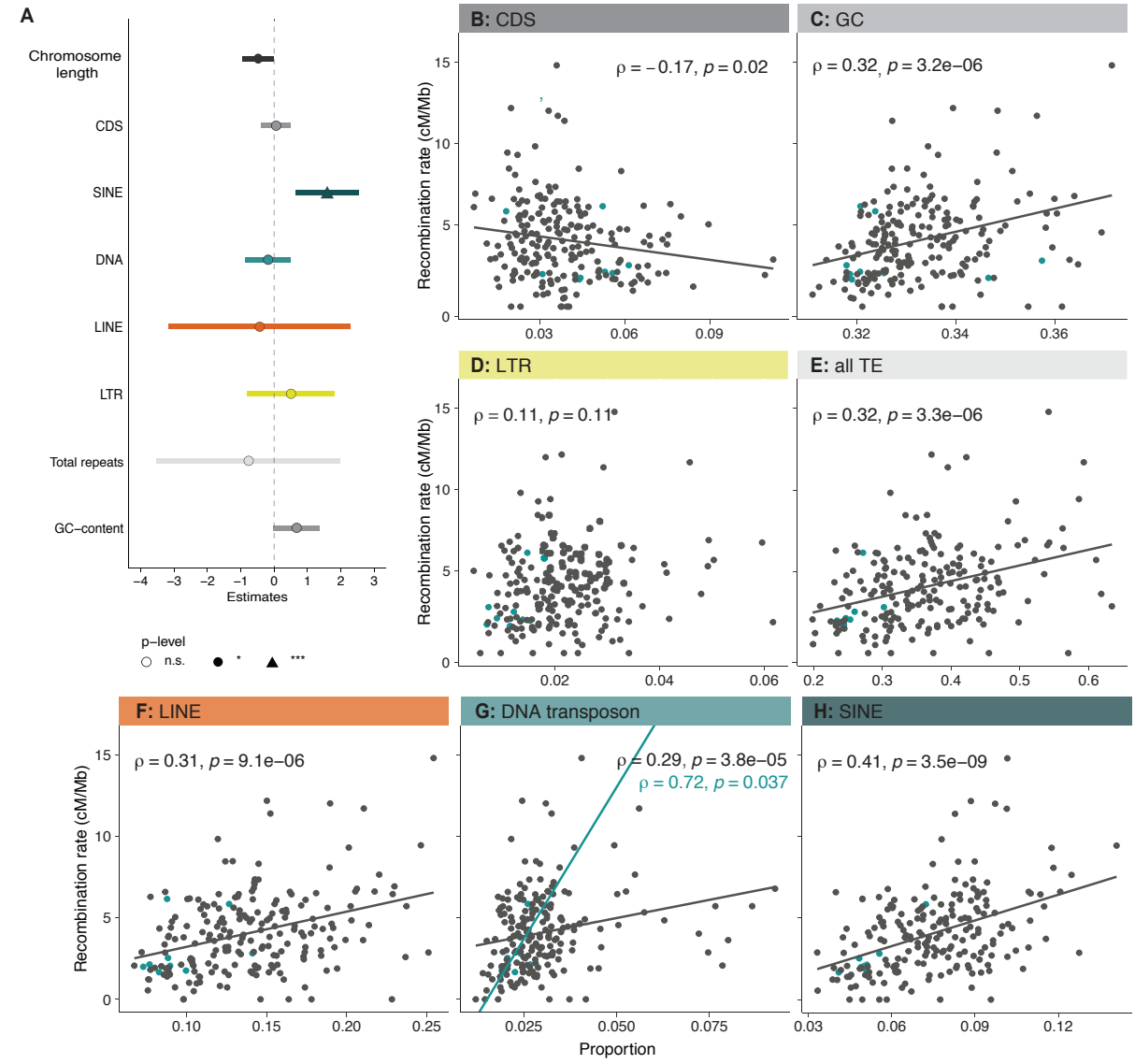

Figure 5 Correlation between recombination rate and density of genomic features. A) Summary of the linear model with regional recombination rate as the response variable. Each explanatory variable in the model is listed along the Y-axis and the relative estimated effect (X-axis), and error intervals 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

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.

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 (Mita et al., 2004). 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 (Lohse et al., 2021), 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 (Mita et al., 2004). 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 (Bachtrog, 2013). 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 (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 > 300 different larval host-plants in 11 plant families (Ackery, 1988; Celorio-Mancera et al., 2016; 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., 2016). 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 the 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 increased recombination rate away from the center and significant drops 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, 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 (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. 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 (Deininger and Batzer, 1999). 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 (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 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 (Aljanabi, 1997). 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 (Andrews et al., 2012). The reads were filtered using the Stacks2 modules `clone_filter` to remove PCR-duplicates and `process_radtags` to filter for quality. We evaluated phredscore in sliding windows covering 15% of the read length and removed reads with mean score below 10 (Catchen et al., 2013). 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 (Lohse et al., 2021) using the bwa mem algorithm (Li, 2013) with default options. Resulting bam files were sorted with samtools sort (Li et al., 2009) 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 (Okonechnikov et al., 2015). 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 (Li et al., 2009). The variants were then converted to likelihoods with Pileup2Likelihoods in LepMap3 using default settings (Rastas, 2017). The LepMap3 protocol (Rastas, 2017) 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 (Gurevich et al., 2013). For the subsequent analysis we excluded unassembled haplotigs from the genome assembly and retained all the other scaffolds.

We used MCScanX (Wang et al., 2012) 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 (Krzywinski et al., 2009) 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 (Holt and Yandell, 2011) iteratively in three steps. In the first step, we mapped previously available transcriptomic evidence data from the painted lady based on wing tissue (Connahs et al., 2016)(accessed on 2020-05-15) and masked all known repeats. RepeatMasker version 4.0.3 (Smit et al., 2015) was used within the MAKER pipeline with a manually curated Lepidoptera repeat library (Talla et al., 2017) 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 (Korf, 2004). 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 (Jones et al., 2014). 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 (Bao et al., 2015) and curated repeats from the monarch butterfly (Zhan and Reppert, 2013). RepeatModeler and RepeatMasker (Holt and Yandell, 2011) were thereafter run again with the updated repeat database.

Coordinates of newly identified repeats were intersected with gene model positions using BEDTools (Quinlan and Hall, 2010) 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 (Huerta-Cepas et al., 2019). The entire painted lady gene set represented 89.9% of the complete BUSCO arthropod gene set (Manni et al., 2021).

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 (Emms and Kelly, 2019) 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 (Librado et al., 2012). We used the species tree obtained with OrthoFinder as input, with additional conversions using Tree as implemented in ete3 (Huerta-Cepas et al., 2016).

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) (Akaike, 1974), 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 (Wolfe et al., 2013), 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 (Alexa and Rahnenfuhrer, 2021) in R version 4.1.0 (R Core Team (2021), 2021). 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 (Alexa et al., 2006). 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 (Benjamini and Hochberg, 1995).

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 (Rezvoy et al., 2007).

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 gained gene. Potential differences in densities of genomic features between bins were assessed with two-sided Wilcoxon rank sum tests in R (Bauer, 1972).

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 (Best and Roberts, 1975), after testing for deviation from normal distribution with the Shapiro-Wilk normality test (Royston, 1982). 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 (Wickham, 2009).

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 contribution statement

DS: Conceptualization, Formal Analysis, Investigation, Data Curation, Writing, Visualization
KN: Conceptualization, Formal Analysis, Investigation, Data Curation, Writing, Visualization
LH: Conceptualization, Investigation, Writing - Review & Editing
RV: Conceptualization, Writing - Review & Editing, Funding acquisition
GT: Conceptualization, Writing - Review & Editing, Funding acquisition
NB: Conceptualization, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

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.

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

Abe, H., Fujii, T., Tanaka, N., Yokoyama, T., Kakehashi, H., Ajimura, M., Mita, K., Banno, Y., Yasukochi, Y., Oshiki, T., Nenoi, M., Ishikawa, T. and Shimada, T. (2008), 'Identification of the female-determining region of the W chromosome in *Bombyx mori*', *Genetica* **133**(3), 269–282.

Ackery, P. R. (1988), 'Hostplants and classification: a review of nymphalid butterflies', *Biological Journal of the Linnean Society* **33**(2), 95–203.

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. and Hanski, I. (2014), 'The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera', *Nature Communications* **5**(1), 4737.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Alexa, A. and Rahnenfuhrer, J. (2021), 'topGO: enrichment analysis for gene ontology'.

Alexa, A., Rahnenführer, J. and Lengauer, T. (2006), 'Improved scoring of functional groups from gene expression data by decorrelating GO graph structure', *Bioinformatics* **22**(13), 1600–1607.

Aljanabi, S. (1997), 'Universal and rapid salt-extraction of high quality genomic DNA for PCR- based techniques', *Nucleic Acids Research* **25**(22), 4692–4693.

Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C. and Wingett, S. (2012), 'FastQC', Babraham Institute.

Apuli, R.-P., Bernhardsson, C., Schiffthaler, B., Robinson, K. M., Jansson, S., Street, N. R. and Ingvarsson, P. K. (2020), 'Inferring the genomic landscape of recombination rate variation in european aspen (*Populus tremula*)', *G3* **10**(1), 299–309.

Bachtrog, D. (2013), 'Y chromosome evolution: emerging insights into processes of Y chromosome degeneration', *Nature reviews. Genetics* **14**(2), 113–124.

Bao, W., Kojima, K. K. and Kohany, O. (2015), 'Repbase update, a database of repetitive elements in eukaryotic genomes', *Mobile DNA* **6**(1), 11.

Barnes, T. M., Kohara, Y., Coulson, A. and Hekimi, S. (1995), 'Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*', *Genetics* **141**(1), 159–179.

Bauer, D. F. (1972), 'Constructing confidence sets using rank statistics', *Journal of the American Statistical Association* **67**(339), 687–690.

Beldade, P., Saenko, S. V., Pul, N. and Long, A. D. (2009), '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), e1000366.

Benjamini, Y. and Hochberg, Y. (1995), '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.

Best, D. J. and Roberts, D. E. (1975), 'Algorithm AS 89: The upper tail probabilities of Spearman's rho', *Applied Statistics* **24**(3), 377.

Boman, J., Mugal, C. F. and Backström, N. (2021), 'The effects of GC-biased gene conversion on patterns of genetic diversity among and across butterfly genomes', *Genome Biology and Evolution* **13**(5), evab064.

Breeschoten, T., van der Linden, C. F. H., Ros, V. I. D., Schranz, M. E. and Simon, S. (2022), 'Expanding the menu: are polyphagy and gene family expansions linked across Lepidoptera?', *Genome Biology and Evolution* **14**(1), evab283.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A. and Cresko, W. A. (2013), 'Stacks: An analysis tool set for population genomics', *Molecular Ecology* **22**(11), 3124–3140.

Celorio-Mancera, M. d. I. P., Rastas, P., Steward, R. A., Nylin, S. and Wheat, C. W. (2021), 'Chromosome level assembly of the comma butterfly (*Polygonia c-album*)', *Genome Biology and Evolution* .

Celorio-Mancera, M. d. I. P., Wheat, C. W., Huss, M., Vezzi, F., Neethiraj, R., Reimegård, J., Nylin, S. and Janz, N. (2016), 'Evolutionary history of host use, rather than plant phylogeny, determines gene expression in a generalist butterfly', *BMC Evolutionary Biology* (1), 59.

Challi, R. J., Kumar, S., Dasmahapatra, K. K., Jiggins, C. D. and Blaxter, M. (2016), 'Lepbase: The Lepidopteran genome database', p. 056994.

Chapman, J. W., Reynolds, D. R. and Wilson, K. (2015), 'Long-range seasonal migration in insects: mechanisms, evolutionary drivers and ecological consequences', *Ecology Letters* **18**(3), 287–302.

Chen, S., Krinsky, B. H. and Long, M. (2013), 'New genes as drivers of phenotypic evolution', *Nature Reviews Genetics* **14**(9), 645–660.

Chowdhury, S., Fuller, R. A., Dingle, H., Chapman, J. W. and Zalucki, M. P. (2021), 'Migration in butterflies: a global overview', *Biological Reviews* **96**(4), 1462–1483.

Connahs, H., Rhen, T. and Simmons, R. B. (2016), 'Transcriptome analysis of the painted lady butterfly, *Vanessa cardui* during wing color pattern development', *BMC Genomics* **17**(1), 270.

Davey, J. W., Barker, S. L., Rastas, P. M., Pinharanda, A., Martin, S. H., Durbin, R., McMillan, W. O., Merrill, R. M. and Jiggins, C. D. (2017), 'No evidence for maintenance of a sympatric *Heliconius* species barrier by chromosomal inversions', *Evolution Letters* **1**(3), 138–154.

de Vos, J. M., Augustijnen, H., Bätscher, L. and Lucek, K. (2020), 'Speciation through chromosomal fusion and fission in Lepidoptera', *Philosophical Transactions of the Royal Society B: Biological Sciences* **375**(1806), 20190539.

Deininger, P. L. and Batzer, M. A. (1999), 'Alu repeats and human disease', *Molecular Genetics and Metabolism* **67**(3), 183–193.

Emerson, J. J., Cardoso-Moreira, M., Borevitz, J. O. and Long, M. (2008), 'Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*', *Science* **320**(5883), 1629–1631.

Emms, D. M. and Kelly, S. (2019), 'OrthoFinder: Phylogenetic orthology inference for comparative genomics', *Genome Biology* **20**(1), 238.

Fraïsse, C., Picard, M. A. L. and Vicoso, B. (2017), 'The deep conservation of the Lepidoptera Z chromosome suggests a non-canonical origin of the W', *Nature Communications* **8**(1), 1486.

Garcìa-Berro, A., Talla, V., Vila, R., Wai, H. K., Shipilina, D., Chan, K. G., Pierce, N. E., Backström, N. and Talavera, G. (in prep), 'Genomic demographic inference shows migratory butterflies display higher heterozygosity and long-term effective population size'.

Gu, L., Reilly, P. F., Lewis, J. J., Reed, R. D., Andolfatto, P. and Walters, J. R. (2019), 'Dichotomy of dosage compensation along the neo z-chromosome of the monarch butterfly', *Current Biology* **29**(23), 4071–4077.e3.

Guerra, P. A., Gegear, R. J. and Reppert, S. M. (2014), 'A magnetic compass aids monarch butterfly migration', *Nature Communications* **5**(1), 4164.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G. (2013), 'QUAST: Quality assessment tool for genome assemblies', *Bioinformatics* **29**(8), 1072–1075.

Haenel, Q., Laurentino, T. G., Roesti, M. and Berner, D. (2018), 'Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics', *Molecular Ecology* **27**(11), 2477–2497.

Hara, S. and Yamakawa, M. (1995), 'Moricin, a novel type of antibacterial peptide isolated from the silkworm, *Bombyx mori*', *The Journal of Biological Chemistry* **270**(50), 29923–29927.

Hatfield, M. J., Umans, R. A., Hyatt, J. L., Edwards, C. C., Wierdl, M., Tsurkan, L., Taylor, M. R. and Potter, P. M. (2016), 'Carboxylesterases: general detoxifying enzymes', *Chemico-biological interactions* **259**, 327–331.

Hedges, D. and Deininger, P. (2007), 'Inviting instability: transposable elements, double-strand breaks, and the maintenance of genome integrity', *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **616**(1-2), 46–59.

Henikoff, S. (1997), 'Gene families: the taxonomy of protein paralogs and chimeras', *Science* **278**(5338), 609–614.

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. and Wheat, C. W. (2019), 'Unprecedented reorganization of holocentric chromosomes provides insights into the enigma of lepidopteran chromosome evolution', *Science Advances* **5**(6), eaau3648.

Holt, C. and Yandell, M. (2011), 'MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects', *BMC Bioinformatics* **12**(1), 491.

Huang, F.-F., Chai, C.-L., Zhang, Z., Liu, Z.-H., Dai, F.-Y., Lu, C. and Xiang, Z.-H. (2008), 'The UDP-glucosyltransferase multigene family in *Bombyx mori*', *BMC Genomics* **9**, 563.

Huerta-Cepas, J., Serra, F. and Bork, P. (2016), 'ETE 3: reconstruction, analysis, and visualization of phylogenomic data', *Molecular Biology and Evolution* **33**(6), 1635–1638.

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. and Bork, P. (2019), 'eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses', *Nucleic Acids Research* **47**(D1), D309–D314.

Jiggins, C. D., Mavarez, J., Beltrán, M., McMillan, W. O., Johnston, J. S. and Bermingham, E. (2005), 'A genetic linkage map of the mimetic butterfly *Heliconius melpomene*', *Genetics* **171**(2), 557–570.

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. and Hunter, S. (2014), 'InterProScan 5: genome-scale protein function classification', *Bioinformatics* **30**(9), 1236–1240.

Kang, L., Chen, X., Zhou, Y., Liu, B., Zheng, W., Li, R., Wang, J. and Yu, J. (2004), '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.

Kawakami, T., Mugal, C. F., Suh, A., Nater, A., Burri, R., Smeds, L. and Ellegren, H. (2017), '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.

Kawakami, T., Smeds, L., Backström, N., Husby, A., Qvarnström, A., Mugal, C. F., Olason, P. and Ellegren, H. (2014), '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.

Kazazian, H. H. (2004), 'Mobile elements: drivers of genome evolution', *Science* **303**(5664), 1626–1632.

Kent, T. V., Uzunović, J. and Wright, S. I. (2017), 'Coevolution between transposable elements and recombination', *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**(1736), 20160458.

Kidwell, M. G. and Lisch, D. (1997), 'Transposable elements as sources of variation in animals and plants', *Proceedings of the National Academy of Sciences* **94**(15), 7704–7711.

Kondrashov, F. A. (2012), 'Gene duplication as a mechanism of genomic adaptation to a changing environment', *Proceedings of the Royal Society B: Biological Sciences* **279**(1749), 5048–5057.

Korf, I. (2004), 'Gene finding in novel genomes', *BMC Bioinformatics* **5**(1), 59.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. and Marra, M. A. (2009), 'Circos: An information aesthetic for comparative genomics', *Genome Research* **19**(9), 1639–1645.

Landys, M. M., Piersma, T., Guglielmo, C. G., Jukema, J., Ramenofsky, M. and Wingfield, J. C. (2005), 'Metabolic profile of long-distance migratory flight and stopover in a shorebird', *Proceedings of the Royal Society B: Biological Sciences* **272**(1560), 295–302.

Lewis, J. J., Cicconardi, F., Martin, S. H., Reed, R. D., Danko, C. G. and Montgomery, S. H. (2021), 'The *Dryas iulia* genome supports multiple gains of a W chromosome from a B chromosome in butterflies', *Genome Biology and Evolution* **13**(7).

Li, F., Zhao, X., Li, M., He, K., Huang, C., Zhou, Y., Li, Z. and Walters, J. R. (2019), 'Insect genomes: progress and challenges', *Insect Molecular Biology* **28**(6), 739–758.

Li, H. (2013), 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM'.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup (2009), 'The sequence alignment/map format and SAMtools', *Bioinformatics (Oxford, England)* **25**(16), 2078–2079.

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. and Wang, W. (2015), 'Outbred genome sequencing and crispr/cas9 gene editing in butterflies', *Nature Communications* **6**(1), 8212.

Librado, P., Vieira, F. G. and Rozas, J. (2012), 'BadiRate: estimating family turnover rates by likelihood-based methods', *Bioinformatics* **28**(2), 279–281.

Lipinski, K. J., Farslow, J. C., Fitzpatrick, K. A., Lynch, M., Katju, V. and Bergthorsson, U. (2011), 'High spontaneous rate of gene duplication in *Caenorhabditis legans*', *Current biology : CB* **21**(4), 306–310.

Loehlin, D. W. and Carroll, S. B. (2016), 'Expression of tandem gene duplicates is often greater than twofold', *Proceedings of the National Academy of Sciences* **113**(21), 5988–5992.

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 and Darwin Tree of Life Consortium (2021), 'The genome sequence of the painted lady, *Vanessa cardui* linnaeus 1758', *Wellcome Open Research* **6**, 324.

Lukhtanov, V. (2015), '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.

Luque, T., Okano, K. and O'Reilly, D. R. (2002), 'Characterization of a novel silkworm (*Bombyx mori*) phenol UDP-glucosyltransferase', *European Journal of Biochemistry* **269**(3), 819–825.

Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A. and Zdobnov, E. M. (2021), '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.

Martin, S. H., Davey, J. W., Salazar, C. and Jiggins, C. D. (2019), 'Recombination rate variation shapes barriers to introgression across butterfly genomes', *PLOS Biology* **17**(2), e2006288.

McClintock, B. (1956), 'Controlling elements and the gene', *Cold Spring Harbor Symposia on Quantitative Biology* **21**(0), 197–216.

Menchetti, M., Guéguen, M. and Talavera, G. (2019), 'Spatio-temporal ecological niche modelling of multigenerational insect migrations', *Proceedings of the Royal Society B: Biological Sciences* **286**(1910), 20191583.

Merlin, C. and Liedvogel, M. (2019), 'The genetics and epigenetics of animal migration and orientation: birds, butterflies and beyond', *Journal of experimental biology* **222**.

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. and Sasaki, T. (2004), 'The genome sequence of ilkworm, *Bombyx mori*', *DNA Research* **11**(1), 27–35.

Murata, M. and Tojo, S. (2013), 'Utilization of lipid for flight and reproduction in *Spodoptera litura* (Lepidoptera: Noctuidae)', *EJE* **99**(2), 221–224.

Nagare, M., Ayachit, M., Agnihotri, A., Schwab, W. and Joshi, R. (2021), 'Glycosyltransferases: The multifaceted enzymatic regulator in insects', *Insect Molecular Biology* **30**(2), 123–137.

Nylin, S., Slove, J. and Janz, N. (2014), 'Host plant utilization, host range oscilations and diversification in Nymphalid butterflies', *Evolution* **68**(1), 105–124.

Ojeda-López, J., Marczuk-Rojas, J. P., Polushkina, O. A., Purucker, D., Salinas, M. and Carretero-Paulet, L. (2020), 'Evolutionary analysis of the *Moringa oleifera* genome reveals a recent burst of plastid to nucleus gene duplications', *Scientific Reports* **10**(1), 17646.

Okonechnikov, K., Conesa, A. and García-Alcalde, F. (2015), 'Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data', *Bioinformatics* p. btv566.

Peñalba, J. V. and Wolf, J. B. W. (2020), 'From molecules to populations: appreciating and estimating recombination rate variation', *Nature Reviews Genetics* **21**(8), 476–492.

Podsiadlowski, L., Tunström, K., Espeland, M. and Wheat, C. W. (2021), 'The genome assembly and annotation of the apollo butterfly *Parnassius apollo*, a flagship species for conservation biology', *Genome Biology and Evolution* **13**(8), evab122.

Quinlan, A. R. and Hall, I. M. (2010), 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics* **26**(6), 841–842.

R Core Team (2021) (2021), 'R: A language and environment for statistical computing.'.
URL: *http://www.r-project.org/index.html*

Rastas, P. (2017), 'Lep-MAP3: Robust linkage mapping even for low-coverage whole genome sequencing data', *Bioinformatics* **33**(23), 3726–3732.

Ray, D. A., Grimshaw, J. R., Halsey, M. K., Korstian, J. M., Osmanski, A. B., Sullivan, K. A. M., Wolf, K. A., Reddy, H., Foley, N., Stevens, R. D., Knisbacher, B. A., Levy, O., Counterman, B., Edelman, N. B. and Mallet, J. (2019), 'Simultaneous te analysis of 19 heliconiine butterflies yields novel insights into rapid te-based genome diversification and multiple sine births and deaths', *Genome Biology and Evolution* **11**(8), 2162–2177.

Rezvoy, C., Charif, D., Gueguen, L. and Marais, G. A. (2007), 'MareyMap: an R-based tool with graphical interface for estimating recombination rates', *Bioinformatics* **23**(16), 2188–2189.

Rizzon, C., Marais, G., Gouy, M. and Biémont, C. (2002), 'Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome', *Genome Research* **12**(3), 400–407.

Rosser, N., Edelman, N. B., Queste, L. M., Nelson, M., Seixas, F., Dasmahapatra, K. K. and Mallet, J. (2022), 'Complex basis of hybrid female sterility and Haldane's rule in Heliconius butterflies: Z-linkage and epistasis', *Molecular Ecology* (31).

Royston, J. P. (1982), 'An extension of Shapiro and Wilk's W test for normality to large samples', *Applied Statistics* **31**(2), 115.

Scherthan, H., Weich, S., Schwegler, H., Heyting, C., Härle, M. and Cremer, T. (1996), 'Centromere and telomere movements during early meiotic prophase of mouse and man are associated with the onset of chromosome pairing.', *Journal of Cell Biology* **134**(5), 1109–1125.

Schwander, T., Libbrecht, R. and Keller, L. (2014), 'Supergenes and complex phenotypes', *Current Biology* **24**(7), R288–R294.

Shields, O. (1992), 'World distribution of the *Vanessa cardui* group nymphalidae', *Journal of the Lepidopterists' Society* **46**, 235–238.

Smit, A., Hubley, R. and Green, P. (2015), 'RepeatMasker Open-4.0'. **URL:** *http://www.repeatmasker.org*

Smith, G. R. and Nambiar, M. (2020), 'New solutions to old problems: Molecular mechanisms of meiotic crossover control', *Trends in genetics* **36**(5), 337–346.

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. and Saastamoinen, M. (2022), 'Improved chromosome-level genome assembly of the glanville fritillary butterfly (*Melitaea cinxia*) integrating pacific biosciences long reads and a high-density linkage map', *GigaScience* **11**(1), giab097.

Srygley, R. B. and Dudley, R. (2008), 'Optimal strategies for insects migrating in the flight boundary layer: mechanisms and consequences', *Integrative and Comparative Biology* **48**(1), 119–133.

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W. and Smadja, C. M. (2017), '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.

Talavera, G., Bataille, C., Benyamini, D., Gascoigne-Pees, M. and Vila, R. (2018), 'Round-trip across the Sahara: afrotropical painted lady butterflies recolonize the mediterranean in early spring', *Biology Letters* **14**(6), 20180274.

Talavera, G. and Vila, R. (2016), '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* **120** (2), 274–285.

Talla, V., Soler, L., Kawakami, T., Dincă, V., Vila, R., Friberg, M., Wiklund, C. and Backström, N. (2019), '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.

Talla, V., Suh, A., Kalsoom, F., Dincă, V., Vila, R., Friberg, M., Wiklund, C. and Backström, N. (2017), '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.

Tiley, G. P. and Burleigh, J. G. (2015), 'The relationship of recombination rate, genome structure, and patterns of molecular evolution across angiosperms', *BMC Evolutionary Biology* **15**(1), 194.

Tobler, A., Kapan, D., Flanagan, N. S., Gonzalez, C., Peterson, E., Jiggins, C. D., Johntson, J. S., Heckel, D. G. and McMillan, W. O. (2005), 'First-generation linkage map of the warningly colored butterfly *Heliconius erato*', *Heredity* **94**(4), 408–417.

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. and Wheat, C. W. (2021), 'A complex interplay between balancing selection and introgression maintains a genus-wide alternative life history strategy', p. 2021.05.20.445023.

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. and Papa, R. (2017), 'Complex modular architecture around a simple toolkit of wing pattern genes', *Nature Ecology & Evolution* **1**(3), 1–12.

Van't Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., Hall, N., Darby, A. C. and Saccheri, I. J. (2016), 'The industrial melanism mutation in british peppered moths is a transposable element', *Nature* **534**(7605), 102–105.

Wahlberg, N. and Rubinoff, D. (2011), '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.

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. and Paterson, A. H. (2012), 'MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity', *Nucleic Acids Research* **40**(7), e49.

Weber, J.-M. (2009), 'The physiology of long-distance migration: extending the limits of endurance metabolism', *Journal of Experimental Biology* **212**(5), 593–597.

Wells, J. N. and Feschotte, C. (2020), 'A field guide to eukaryotic transposable elements', *Annual Review of Genetics* **54**(1), 539–561.

Wickham, H. (2009), *ggplot2: elegant graphics for data analysis*, Use R!, Springer-Verlag, New York.

Wieczorek, H., Beyenbach, K. W., Huss, M. and Vitavska, O. (2009), 'Vacuolar-type proton pumps in insect epithelia', *The Journal of Experimental Biology* **212**(11), 1611–1619.

Wolfe, D., Dudek, S., Ritchie, M. D. and Pendergrass, S. A. (2013), 'Visualizing genomic information across chromosomes with PhenoGram', *BioData Mining* **6**(1), 18.

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. and Mita, K. (2008), 'A BAC-based integrated linkage map of the silkworm *Bombyx mori*', *Genome Biology* **9**(1), R21.

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. and Li, X. (2020), 'Chromosome-level reference genome assembly and gene editing of the dead-leaf butterfly *Kallima inachus*', *Molecular Ecology Resources* **20**(4), 1080–1092.

Yasukochi, Y. (1998), 'A dense genetic map of the silkworm, *Bombyx mori*, covering all chromosomes based on 1018 molecular markers', *Genetics* **150**(4), 1513–1525.

Zhan, S. and Reppert, S. M. (2013), 'MonarchBase: the monarch butterfly genome database', *Nucleic Acids Research* **41**(D1), D758–D763.

Zhan, S., Zhang, W., Niitepõld, K., Hsu, J., Haeger, J. F., Zalucki, M. P., Altizer, S., de Roode, J. C., Reppert, S. M. and Kronforst, M. R. (2014), 'The genetics of monarch butterfly migration and warning colouration', *Nature* **514**(7522), 317–321.

Zhang, L. (2003), 'Does recombination shape the distribution and evolution of tandemly arrayed genes (tags) in the *Arabidopsis thaliana* genome?', *Genome Research* **13**(12), 2533–2540.

Zhang, L., Steward, R. A., Wheat, C. W. and Reed, R. D. (2021), 'High-quality genome assembly and comprehensive transcriptome of the painted lady butterfly *Vanessa cardui*', *Genome Biology and Evolution* **13**(7).

Zhu, H., Gegear, R. J., Casselman, A., Kanginakudru, S. and Reppert, S. M. (2009), 'Defining behavioral and molecular differences between summer and migratory monarch butterflies', *BMC Biology* **7**(1), 14.

Highlights

- Data: Gene and repeat annotations and a recombination map for *Vanessa cardui*.
- Gene families associated with metabolism, detoxification and defence have expanded.
- Chromosome size is a main determinant of the recombination rate in butterflies.
- Recombination is associated with density of transposable elements.
- The W-chromosome is enriched for transposable elements and lacks functional genes.

Click here to access/download
LaTeX Source File
bmcart.cls

Author contribution statement

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

Authors: Daria Shipilina, Karin Näsvall, Lars Höök, Roger Vila, Gerard Talavera and Niclas Backström

Contributions:

DS: Conceptualization, Formal Analysis, Investigation, Data Curation, Writing, Visualization
KN: Conceptualization, Formal Analysis, Investigation, Data Curation, Writing, Visualization
LH: Conceptualization, Investigation, Writing - Review & Editing
RV: Conceptualization, Writing - Review & Editing, Funding acquisition
GT: Conceptualization, Writing - Review & Editing, Funding acquisition
NB: Conceptualization, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

Click here to access/download
Supplementary Material
ShipilinaNasvall_etal_Supplementary.pdf

Revised Manuscript Click here to view linked References

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

Characterization of gene family expansions and crossing over is crucial for understanding how organisms adapt to the environment. Here, we develop a high-density linkage map and detailed genome annotation of the painted lady butterfly (*Vanessa cardui*) - a non-diapausing, highly polyphagous species famous for its long-distance migratory behavior and almost cosmopolitan distribution. Our results reveal a complex interplay between regional recombination rate variation, gene duplications and transposable element activity shaping the genome structure of the painted lady. We identify several lineage specific gene family expansions. Their functions are mainly associated with protein and fat metabolism, detoxification, and defense against infection - critical processes for the painted lady's unique life-history. Furthermore, the detailed recombination maps allow us to characterize the regional recombination landscape, data that reveal a strong effect of chromosome size on the recombination rate, a limited impact of GC-biased gene conversion and a positive association between recombination and short interspersed elements.

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 (Henikoff, 1997; Ojeda-López et al., 2020; Zhang, 2003), and gene family expansions and contractions are important sources for generation of phenotypic diversity (Chen et al., 2013; Kondrashov, 2012). 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 (Schwander et al., 2014).

Since the spearheading work by McClintock (McClintock, 1956), transposable elements (TEs) have been acknowledged as major contributors to different types of evolutionary change in eukaryotes (Kazazian, 2004; Kidwell and Lisch, 1997). Transposable elements are capable of self-replication within the host genome and can mediate both small scale deletions and duplications, large scale chromosomal rearrangements (Kidwell and Lisch, 1997) and considerable genome size expansions (Podsiadlowski et al., 2021; Talla et al., 2017). In 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 (Peñalba and Wolf, 2020). Recombination also influences selection efficiency by directly preventing the accumulation of deleterious alleles (Müller's ratchet) and breaking the physical linkage

*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

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 (Stapley et al., 2017; Tiley and Burleigh, 2015). However, besides detailed recombination maps in the butterfly genus *Heliconius* (Martin et al., 2019), 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 (Haenel et al., 2018; Talla et al., 2019).

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 wide-spread of all butterfly species (Talavera et al., 2018), 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 (Chapman et al., 2015; Guerra et al., 2014). 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 (Menchetti et al., 2019; Talavera and Vila, n.d.). 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 (Garcìa-Berro et al., in prep). This could also be a consequence of the species' ability to utilize a wide range of host plants (Ackery, 1988; Celorio-Mancera et al., 2016). 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 (Merlin and Liedvogel, 2019).

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 (Lohse et al., 2021). 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 (Connahs et al., 2016; Zhang et al., 2021) 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 (Lohse et al., 2021) 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; $R = 0.91$-1.00, p-value $> 1.00 \times 10^{-04}$, 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).

Figure 1 Distribution of repeat classes and genes as estimated along the painted lady chromosomes (100kb 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).

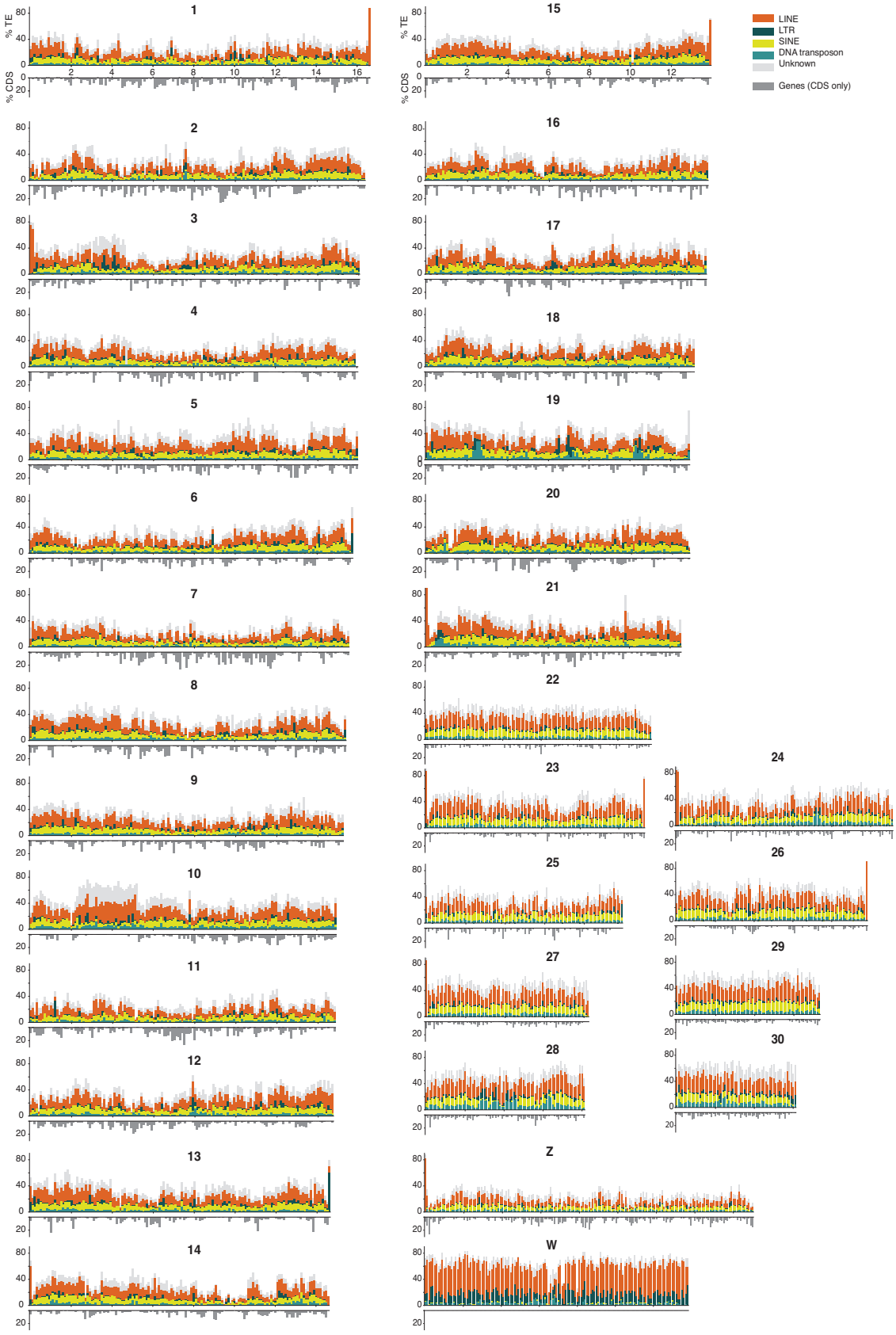

Table 1 Linkage map, genome assembly and annotation statistics

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 (Davey et al., 2017) could also be verified (Figure 2). In summary, this confirms that the painted lady karyotype is highly similar to the inferred ancestral butterfly karyotype (Ahola et al., 2014).

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).

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 (Spearman's rank correlation, $\rho = -0.83$, p-value $= 6.51 \times 10^{-07}$;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 (Figure 4A). Besides

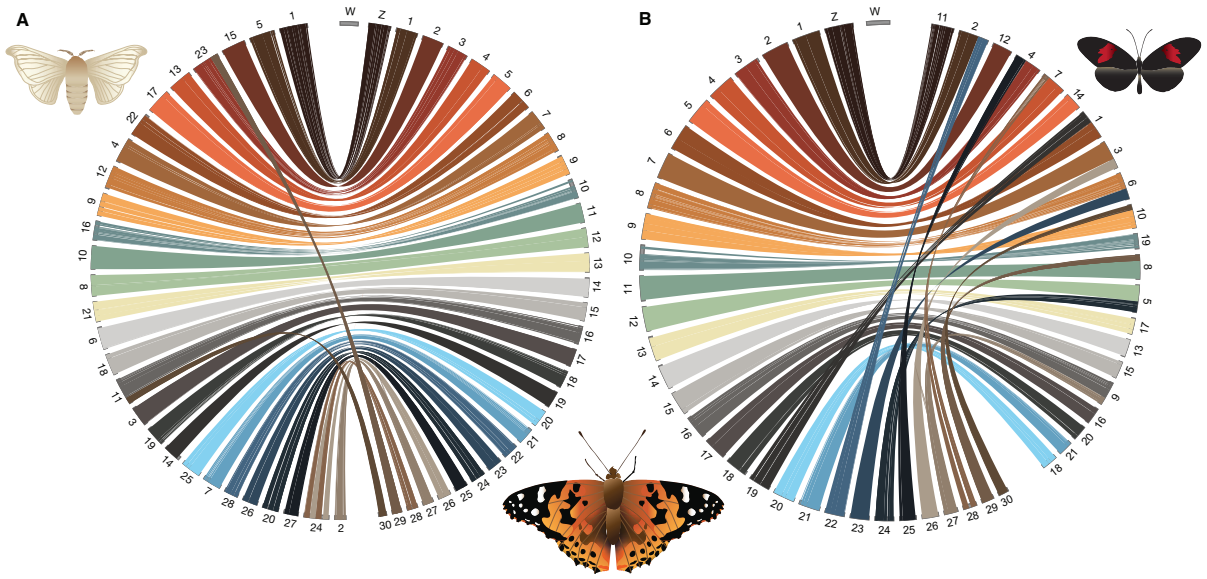

Figure 2 Synteny between the painted lady and a) the silkmoth (*Bombyx mori*) and b) the postman butterfly (*Heliconius melpomene*) chromosomes, respectively. The painted lady autosomes are sorted and named according to length and the sex chromosomes are indicated with Z and W. Orthologous genes are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

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 $= 8.35\times10^{-05}$) and repeat density ($\rho = 0.77$, p-value $= 1.37\times10^{-06}$), and a positive association with gene density ($\rho = 0.68$, p-value $= 2.63\times10^{-05}$)(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 ($\rho = 0.32$, p-value $= 3.68\times10^{-06}$). Gene density was homogeneous across chromosomes, with only a minor increase towards the chromosome center, and was negatively associated with the recombination rate ($\rho = $-0.19, p-value $= 7.27\times10^{-03}$). We found a significant positive association between the overall repeat proportion and the recombination rate ($\rho = 0.35$, p-value $= 3.48\times10^{-07}$; Figure 5), and this pattern was consistent for all repeat classes, but strongest for SINEs ($\rho = 0.42$, p-value $= 2.63\times10^{-10}$) and weakest for LTRs ($\rho = 0.14$, p-value $= 4.04\times10^{-02}$). The association between recombination rate and proportion of LTRs was, however, not significant when only including autosomes ($\rho = 0.11$, p-value $= 11.04\times10^{-02}$; 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

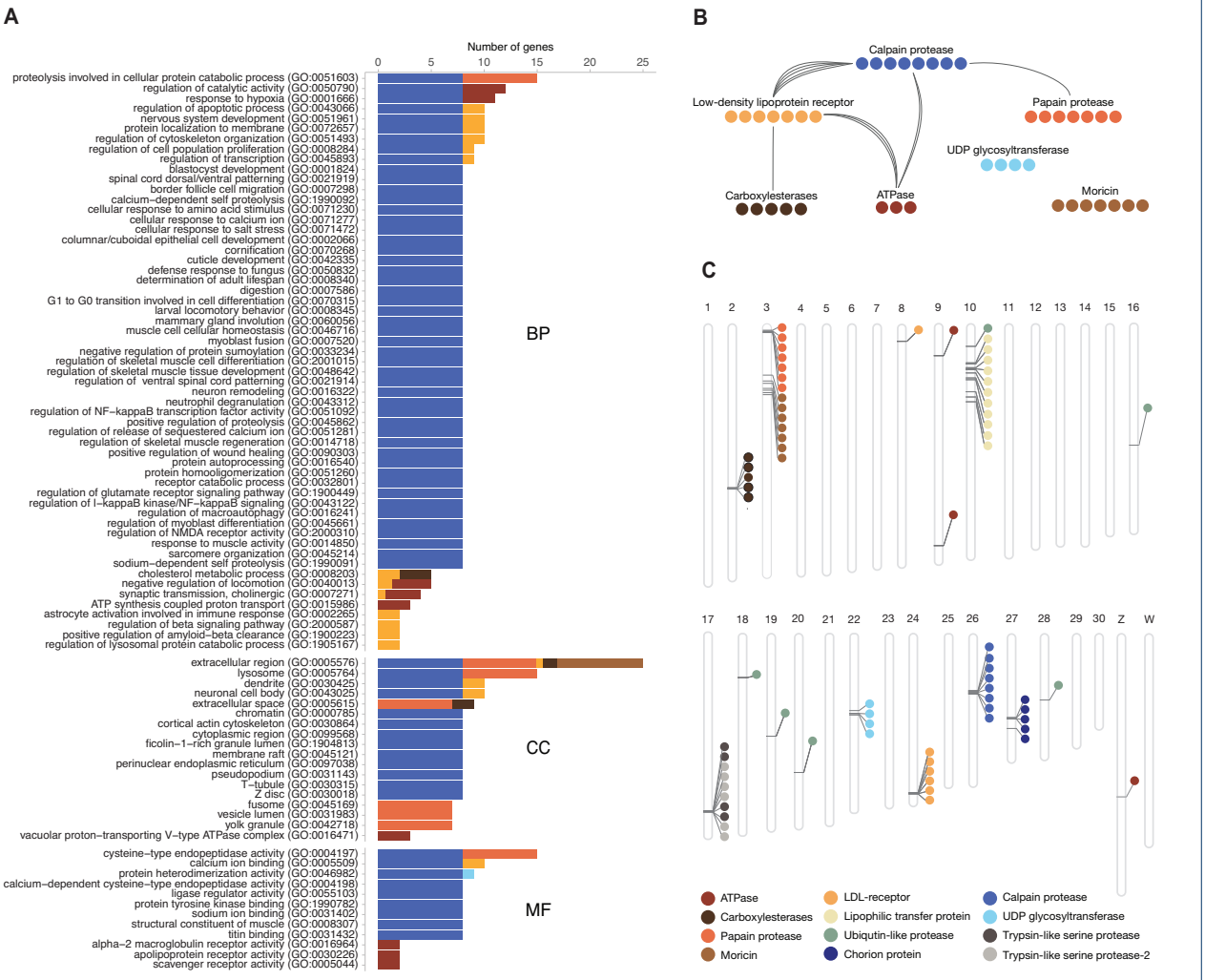

Figure 3 A) Significantly (p-value < 0.05 after FDR-correction) enriched gene ontology (GO) terms associated with expanded gene families in the painted lady. The bars show the number of genes associated with each GO-term. The different GO-categories are biological process (BP), cellular compartment (CC) 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.

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 $= 5.36\times10^{-09}$) and explanatory variables in the model accounted for 21% of the variation in recombination rate ($R2 = 0.24$, $adj\,R2 = 0.21$). Most of the variation was explained by the positive association with the proportion of SINEs (Estimate 1.59, p-value $= 9.75\times10^{-04}$) and the negative association with chromosome size (Estimate -0.48, p-value $= 4.88\times10^{-02}$; 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 $= 9.1\times10^{-01}$ - 8.0×10^{-01}; Figure S6).

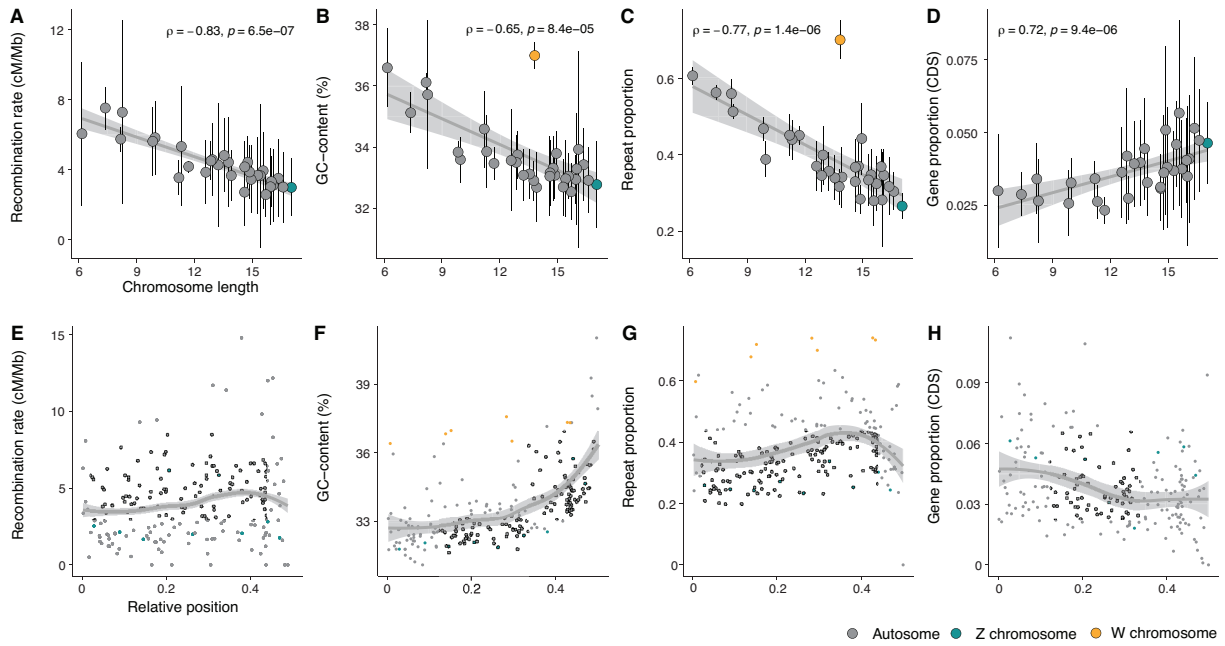

Figure 4 Top row: Associations between chromosome length and A) recombination rate, B) base composition, C) repeat and D) gene proportions. Chromosome length is given in megabases (Mb). Bottom row: Regional distribution of the recombination rate (E), base composition (F), repeat (G) and gene (H) density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

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 (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 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

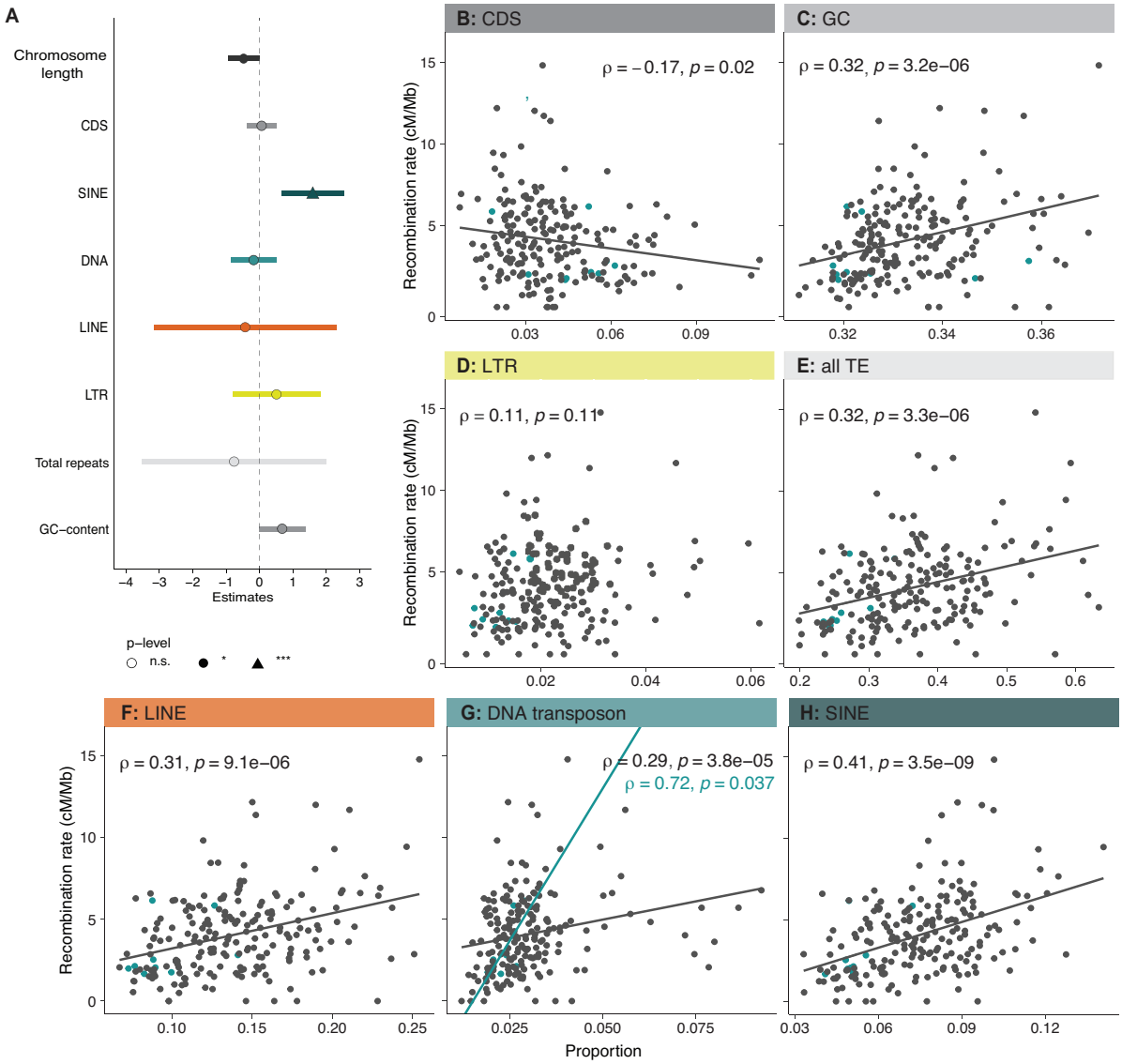

Figure 5 Correlation between recombination rate and density of genomic features. A) Summary of the linear model with regional recombination rate as the response variable. Each explanatory variable in the model is listed along the Y-axis and the relative estimated effect (X-axis), and error intervals 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

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.

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 (Mita et al., 2004). 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 (Lohse et al., 2021), 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 (Mita et al., 2004). 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 (Bachtrog, 2013). 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 (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 > 300 different larval host-plants in 11 plant families (Ackery, 1988; Celorio-Mancera et al., 2016; 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., 2016). 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 the 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 increased recombination rate away from the center and significant drops 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, 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 (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. 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 (Deininger and Batzer, 1999). 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 (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 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 (Aljanabi, 1997). 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 (Andrews et al., 2012). The reads were filtered using the Stacks2 modules `clone_filter` to remove PCR-duplicates and `process_radtags` to filter for quality. We evaluated phredscore in sliding windows covering 15% of the read length and removed reads with mean score below 10 (Catchen et al., 2013). 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 (Lohse et al., 2021) using the bwa mem algorithm (Li, 2013) with default options. Resulting bam files were sorted with samtools sort (Li et al., 2009) 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 (Okonechnikov et al., 2015). 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 (Li et al., 2009). The variants were then converted to likelihoods with Pileup2Likelihoods in LepMap3 using default settings (Rastas, 2017). The LepMap3 protocol (Rastas, 2017) 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 (Gurevich et al., 2013). For the subsequent analysis we excluded unassembled haplotigs from the genome assembly and retained all the other scaffolds.

We used MCScanX (Wang et al., 2012) 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 (Krzywinski et al., 2009) 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 (Holt and Yandell, 2011) iteratively in three steps. In the first step, we mapped previously available transcriptomic evidence data from the painted lady based on wing tissue (Connahs et al., 2016)(accessed on 2020-05-15) and masked all known repeats. RepeatMasker version 4.0.3 (Smit et al., 2015) was used within the MAKER pipeline with a manually curated Lepidoptera repeat library (Talla et al., 2017) 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 (Korf, 2004). 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 (Jones et al., 2014). 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 (Bao et al., 2015) and curated repeats from the monarch butterfly (Zhan and Reppert, 2013). RepeatModeler and RepeatMasker (Holt and Yandell, 2011) were thereafter run again with the updated repeat database.

Coordinates of newly identified repeats were intersected with gene model positions using BEDTools (Quinlan and Hall, 2010) 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 (Huerta-Cepas et al., 2019). The entire painted lady gene set represented 89.9% of the complete BUSCO arthropod gene set (Manni et al., 2021).

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 (Emms and Kelly, 2019) 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 (Librado et al., 2012). We used the species tree obtained with OrthoFinder as input, with additional conversions using Tree as implemented in ete3 (Huerta-Cepas et al., 2016).

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) (Akaike, 1974), 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 (Wolfe et al., 2013), 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 (Alexa and Rahnenfuhrer, 2021) in R version 4.1.0 (R Core Team (2021), 2021). 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 (Alexa et al., 2006). 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 (Benjamini and Hochberg, 1995).

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 (Rezvoy et al., 2007).

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 gained gene. Potential differences in densities of genomic features between bins were assessed with two-sided Wilcoxon rank sum tests in R (Bauer, 1972).

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 (Best and Roberts, 1975), after testing for deviation from normal distribution with the Shapiro-Wilk normality test (Royston, 1982). 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 (Wickham, 2009).

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 contribution statement

DS: Conceptualization, Formal Analysis, Investigation, Data Curation, Writing, Visualization
KN: Conceptualization, Formal Analysis, Investigation, Data Curation, Writing, Visualization
LH: Conceptualization, Investigation, Writing - Review & Editing
RV: Conceptualization, Writing - Review & Editing, Funding acquisition
GT: Conceptualization, Writing - Review & Editing, Funding acquisition
NB: Conceptualization, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

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.

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

Abe, H., Fujii, T., Tanaka, N., Yokoyama, T., Kakehashi, H., Ajimura, M., Mita, K., Banno, Y., Yasukochi, Y., Oshiki, T., Nenoi, M., Ishikawa, T. and Shimada, T. (2008), 'Identification of the female-determining region of the W chromosome in *Bombyx mori*', *Genetica* **133**(3), 269–282.

Ackery, P. R. (1988), 'Hostplants and classification: a review of nymphalid butterflies', *Biological Journal of the Linnean Society* **33**(2), 95–203.

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. and Hanski, I. (2014), 'The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera', *Nature Communications* **5**(1), 4737.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Alexa, A. and Rahnenfuhrer, J. (2021), 'topGO: enrichment analysis for gene ontology'.

Alexa, A., Rahnenführer, J. and Lengauer, T. (2006), 'Improved scoring of functional groups from gene expression data by decorrelating GO graph structure', *Bioinformatics* **22**(13), 1600–1607.

Aljanabi, S. (1997), 'Universal and rapid salt-extraction of high quality genomic DNA for PCR- based techniques', *Nucleic Acids Research* **25**(22), 4692–4693.

Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C. and Wingett, S. (2012), 'FastQC', Babraham Institute.

Apuli, R.-P., Bernhardsson, C., Schiffthaler, B., Robinson, K. M., Jansson, S., Street, N. R. and Ingvarsson, P. K. (2020), 'Inferring the genomic landscape of recombination rate variation in european aspen (*Populus tremula*)', *G3* **10**(1), 299–309.

Bachtrog, D. (2013), 'Y chromosome evolution: emerging insights into processes of Y chromosome degeneration', *Nature reviews. Genetics* **14**(2), 113–124.

Bao, W., Kojima, K. K. and Kohany, O. (2015), 'Repbase update, a database of repetitive elements in eukaryotic genomes', *Mobile DNA* **6**(1), 11.

Barnes, T. M., Kohara, Y., Coulson, A. and Hekimi, S. (1995), 'Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*', *Genetics* **141**(1), 159–179.

Bauer, D. F. (1972), 'Constructing confidence sets using rank statistics', *Journal of the American Statistical Association* **67**(339), 687–690.

Beldade, P., Saenko, S. V., Pul, N. and Long, A. D. (2009), '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), e1000366.

Benjamini, Y. and Hochberg, Y. (1995), '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.

Best, D. J. and Roberts, D. E. (1975), 'Algorithm AS 89: The upper tail probabilities of Spearman's rho', *Applied Statistics* **24**(3), 377.

Boman, J., Mugal, C. F. and Backström, N. (2021), 'The effects of GC-biased gene conversion on patterns of genetic diversity among and across butterfly genomes', *Genome Biology and Evolution* **13**(5), evab064.

Breeschoten, T., van der Linden, C. F. H., Ros, V. I. D., Schranz, M. E. and Simon, S. (2022), 'Expanding the menu: are polyphagy and gene family expansions linked across Lepidoptera?', *Genome Biology and Evolution* **14**(1), evab283.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A. and Cresko, W. A. (2013), 'Stacks: An analysis tool set for population genomics', *Molecular Ecology* **22**(11), 3124–3140.

Celorio-Mancera, M. d. I. P., Rastas, P., Steward, R. A., Nylin, S. and Wheat, C. W. (2021), 'Chromosome level assembly of the comma butterfly (*Polygonia c-album*)', *Genome Biology and Evolution* .

Celorio-Mancera, M. d. I. P., Wheat, C. W., Huss, M., Vezzi, F., Neethiraj, R., Reimegård, J., Nylin, S. and Janz, N. (2016), 'Evolutionary history of host use, rather than plant phylogeny, determines gene expression in a generalist butterfly', *BMC Evolutionary Biology* (1), 59.

Challi, R. J., Kumar, S., Dasmahapatra, K. K., Jiggins, C. D. and Blaxter, M. (2016), 'Lepbase: The Lepidopteran genome database', p. 056994.

Chapman, J. W., Reynolds, D. R. and Wilson, K. (2015), 'Long-range seasonal migration in insects: mechanisms, evolutionary drivers and ecological consequences', *Ecology Letters* **18**(3), 287–302.

Chen, S., Krinsky, B. H. and Long, M. (2013), 'New genes as drivers of phenotypic evolution', *Nature Reviews Genetics* **14**(9), 645–660.

Chowdhury, S., Fuller, R. A., Dingle, H., Chapman, J. W. and Zalucki, M. P. (2021), 'Migration in butterflies: a global overview', *Biological Reviews* **96**(4), 1462–1483.

Connahs, H., Rhen, T. and Simmons, R. B. (2016), 'Transcriptome analysis of the painted lady butterfly, *Vanessa cardui* during wing color pattern development', *BMC Genomics* **17**(1), 270.

Davey, J. W., Barker, S. L., Rastas, P. M., Pinharanda, A., Martin, S. H., Durbin, R., McMillan, W. O., Merrill, R. M. and Jiggins, C. D. (2017), 'No evidence for maintenance of a sympatric *Heliconius* species barrier by chromosomal inversions', *Evolution Letters* **1**(3), 138–154.

de Vos, J. M., Augustijnen, H., Bätscher, L. and Lucek, K. (2020), 'Speciation through chromosomal fusion and fission in Lepidoptera', *Philosophical Transactions of the Royal Society B: Biological Sciences* **375**(1806), 20190539.

Deininger, P. L. and Batzer, M. A. (1999), 'Alu repeats and human disease', *Molecular Genetics and Metabolism* **67**(3), 183–193.

Emerson, J. J., Cardoso-Moreira, M., Borevitz, J. O. and Long, M. (2008), 'Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*', *Science* **320**(5883), 1629–1631.

Emms, D. M. and Kelly, S. (2019), 'OrthoFinder: Phylogenetic orthology inference for comparative genomics', *Genome Biology* **20**(1), 238.

Fraïsse, C., Picard, M. A. L. and Vicoso, B. (2017), 'The deep conservation of the Lepidoptera Z chromosome suggests a non-canonical origin of the W', *Nature Communications* **8**(1), 1486.

Garcìa-Berro, A., Talla, V., Vila, R., Wai, H. K., Shipilina, D., Chan, K. G., Pierce, N. E., Backström, N. and Talavera, G. (in prep), 'Genomic demographic inference shows migratory butterflies display higher heterozygosity and long-term effective population size'.

Gu, L., Reilly, P. F., Lewis, J. J., Reed, R. D., Andolfatto, P. and Walters, J. R. (2019), 'Dichotomy of dosage compensation along the neo z-chromosome of the monarch butterfly', *Current Biology* **29**(23), 4071–4077.e3.

Guerra, P. A., Gegear, R. J. and Reppert, S. M. (2014), 'A magnetic compass aids monarch butterfly migration', *Nature Communications* **5**(1), 4164.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G. (2013), 'QUAST: Quality assessment tool for genome assemblies', *Bioinformatics* **29**(8), 1072–1075.

Haenel, Q., Laurentino, T. G., Roesti, M. and Berner, D. (2018), 'Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics', *Molecular Ecology* **27**(11), 2477–2497.

Hara, S. and Yamakawa, M. (1995), 'Moricin, a novel type of antibacterial peptide isolated from the silkworm, *Bombyx mori*', *The Journal of Biological Chemistry* **270**(50), 29923–29927.

Hatfield, M. J., Umans, R. A., Hyatt, J. L., Edwards, C. C., Wierdl, M., Tsurkan, L., Taylor, M. R. and Potter, P. M. (2016), 'Carboxylesterases: general detoxifying enzymes', *Chemico-biological interactions* **259**, 327–331.

Hedges, D. and Deininger, P. (2007), 'Inviting instability: transposable elements, double-strand breaks, and the maintenance of genome integrity', *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **616**(1-2), 46–59.

Henikoff, S. (1997), 'Gene families: the taxonomy of protein paralogs and chimeras', *Science* **278**(5338), 609–614.

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. and Wheat, C. W. (2019), 'Unprecedented reorganization of holocentric chromosomes provides insights into the enigma of lepidopteran chromosome evolution', *Science Advances* **5**(6), eaau3648.

Holt, C. and Yandell, M. (2011), 'MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects', *BMC Bioinformatics* **12**(1), 491.

Huang, F.-F., Chai, C.-L., Zhang, Z., Liu, Z.-H., Dai, F.-Y., Lu, C. and Xiang, Z.-H. (2008), 'The UDP-glucosyltransferase multigene family in *Bombyx mori*', *BMC Genomics* **9**, 563.

Huerta-Cepas, J., Serra, F. and Bork, P. (2016), 'ETE 3: reconstruction, analysis, and visualization of phylogenomic data', *Molecular Biology and Evolution* **33**(6), 1635–1638.

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. and Bork, P. (2019), 'eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses', *Nucleic Acids Research* **47**(D1), D309–D314.

Jiggins, C. D., Mavarez, J., Beltrán, M., McMillan, W. O., Johnston, J. S. and Bermingham, E. (2005), 'A genetic linkage map of the mimetic butterfly *Heliconius melpomene*', *Genetics* **171**(2), 557–570.

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. and Hunter, S. (2014), 'InterProScan 5: genome-scale protein function classification', *Bioinformatics* **30**(9), 1236–1240.

Kang, L., Chen, X., Zhou, Y., Liu, B., Zheng, W., Li, R., Wang, J. and Yu, J. (2004), '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.

Kawakami, T., Mugal, C. F., Suh, A., Nater, A., Burri, R., Smeds, L. and Ellegren, H. (2017), '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.

Kawakami, T., Smeds, L., Backström, N., Husby, A., Qvarnström, A., Mugal, C. F., Olason, P. and Ellegren, H. (2014), '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.

Kazazian, H. H. (2004), 'Mobile elements: drivers of genome evolution', *Science* **303**(5664), 1626–1632.

Kent, T. V., Uzunović, J. and Wright, S. I. (2017), 'Coevolution between transposable elements and recombination', *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**(1736), 20160458.

Kidwell, M. G. and Lisch, D. (1997), 'Transposable elements as sources of variation in animals and plants', *Proceedings of the National Academy of Sciences* **94**(15), 7704–7711.

Kondrashov, F. A. (2012), 'Gene duplication as a mechanism of genomic adaptation to a changing environment', *Proceedings of the Royal Society B: Biological Sciences* **279**(1749), 5048–5057.

Korf, I. (2004), 'Gene finding in novel genomes', *BMC Bioinformatics* **5**(1), 59.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. and Marra, M. A. (2009), 'Circos: An information aesthetic for comparative genomics', *Genome Research* **19**(9), 1639–1645.

Landys, M. M., Piersma, T., Guglielmo, C. G., Jukema, J., Ramenofsky, M. and Wingfield, J. C. (2005), 'Metabolic profile of long-distance migratory flight and stopover in a shorebird', *Proceedings of the Royal Society B: Biological Sciences* **272**(1560), 295–302.

Lewis, J. J., Cicconardi, F., Martin, S. H., Reed, R. D., Danko, C. G. and Montgomery, S. H. (2021), 'The *Dryas iulia* genome supports multiple gains of a W chromosome from a B chromosome in butterflies', *Genome Biology and Evolution* **13**(7).

Li, F., Zhao, X., Li, M., He, K., Huang, C., Zhou, Y., Li, Z. and Walters, J. R. (2019), 'Insect genomes: progress and challenges', *Insect Molecular Biology* **28**(6), 739–758.

Li, H. (2013), 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM'.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup (2009), 'The sequence alignment/map format and SAMtools', *Bioinformatics (Oxford, England)* **25**(16), 2078–2079.

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. and Wang, W. (2015), 'Outbred genome sequencing and crispr/cas9 gene editing in butterflies', *Nature Communications* **6**(1), 8212.

Librado, P., Vieira, F. G. and Rozas, J. (2012), 'BadiRate: estimating family turnover rates by likelihood-based methods', *Bioinformatics* **28**(2), 279–281.

Lipinski, K. J., Farslow, J. C., Fitzpatrick, K. A., Lynch, M., Katju, V. and Bergthorsson, U. (2011), 'High spontaneous rate of gene duplication in *Caenorhabditis legans*', *Current biology : CB* **21**(4), 306–310.

Loehlin, D. W. and Carroll, S. B. (2016), 'Expression of tandem gene duplicates is often greater than twofold', *Proceedings of the National Academy of Sciences* **113**(21), 5988–5992.

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 and Darwin Tree of Life Consortium (2021), 'The genome sequence of the painted lady, *Vanessa cardui* linnaeus 1758', *Wellcome Open Research* **6**, 324.

Lukhtanov, V. (2015), '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.

Luque, T., Okano, K. and O'Reilly, D. R. (2002), 'Characterization of a novel silkworm (*Bombyx mori*) phenol UDP-glucosyltransferase', *European Journal of Biochemistry* **269**(3), 819–825.

Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A. and Zdobnov, E. M. (2021), '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.

Martin, S. H., Davey, J. W., Salazar, C. and Jiggins, C. D. (2019), 'Recombination rate variation shapes barriers to introgression across butterfly genomes', *PLOS Biology* **17**(2), e2006288.

McClintock, B. (1956), 'Controlling elements and the gene', *Cold Spring Harbor Symposia on Quantitative Biology* **21**(0), 197–216.

Menchetti, M., Guéguen, M. and Talavera, G. (2019), 'Spatio-temporal ecological niche modelling of multigenerational insect migrations', *Proceedings of the Royal Society B: Biological Sciences* **286**(1910), 20191583.

Merlin, C. and Liedvogel, M. (2019), 'The genetics and epigenetics of animal migration and orientation: birds, butterflies and beyond', *Journal of experimental biology* **222**.

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. and Sasaki, T. (2004), 'The genome sequence of ilkworm, *Bombyx mori*', *DNA Research* **11**(1), 27–35.

Murata, M. and Tojo, S. (2013), 'Utilization of lipid for flight and reproduction in *Spodoptera litura* (Lepidoptera: Noctuidae)', *EJE* **99**(2), 221–224.

Nagare, M., Ayachit, M., Agnihotri, A., Schwab, W. and Joshi, R. (2021), 'Glycosyltransferases: The multifaceted enzymatic regulator in insects', *Insect Molecular Biology* **30**(2), 123–137.

Nylin, S., Slove, J. and Janz, N. (2014), 'Host plant utilization, host range oscilations and diversification in Nymphalid butterflies', *Evolution* **68**(1), 105–124.

Ojeda-López, J., Marczuk-Rojas, J. P., Polushkina, O. A., Purucker, D., Salinas, M. and Carretero-Paulet, L. (2020), 'Evolutionary analysis of the *Moringa oleifera* genome reveals a recent burst of plastid to nucleus gene duplications', *Scientific Reports* **10**(1), 17646.

Okonechnikov, K., Conesa, A. and García-Alcalde, F. (2015), 'Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data', *Bioinformatics* p. btv566.

Peñalba, J. V. and Wolf, J. B. W. (2020), 'From molecules to populations: appreciating and estimating recombination rate variation', *Nature Reviews Genetics* **21**(8), 476–492.

Podsiadlowski, L., Tunström, K., Espeland, M. and Wheat, C. W. (2021), 'The genome assembly and annotation of the apollo butterfly *Parnassius apollo*, a flagship species for conservation biology', *Genome Biology and Evolution* **13**(8), evab122.

Quinlan, A. R. and Hall, I. M. (2010), 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics* **26**(6), 841–842.

R Core Team (2021) (2021), 'R: A language and environment for statistical computing.'.
URL: *http://www.r-project.org/index.html*

Rastas, P. (2017), 'Lep-MAP3: Robust linkage mapping even for low-coverage whole genome sequencing data', *Bioinformatics* **33**(23), 3726–3732.

Ray, D. A., Grimshaw, J. R., Halsey, M. K., Korstian, J. M., Osmanski, A. B., Sullivan, K. A. M., Wolf, K. A., Reddy, H., Foley, N., Stevens, R. D., Knisbacher, B. A., Levy, O., Counterman, B., Edelman, N. B. and Mallet, J. (2019), 'Simultaneous te analysis of 19 heliconiine butterflies yields novel insights into rapid te-based genome diversification and multiple sine births and deaths', *Genome Biology and Evolution* **11**(8), 2162–2177.

Rezvoy, C., Charif, D., Gueguen, L. and Marais, G. A. (2007), 'MareyMap: an R-based tool with graphical interface for estimating recombination rates', *Bioinformatics* **23**(16), 2188–2189.

Rizzon, C., Marais, G., Gouy, M. and Biémont, C. (2002), 'Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome', *Genome Research* **12**(3), 400–407.

Rosser, N., Edelman, N. B., Queste, L. M., Nelson, M., Seixas, F., Dasmahapatra, K. K. and Mallet, J. (2022), 'Complex basis of hybrid female sterility and Haldane's rule in Heliconius butterflies: Z-linkage and epistasis', *Molecular Ecology* (31).

Royston, J. P. (1982), 'An extension of Shapiro and Wilk's W test for normality to large samples', *Applied Statistics* **31**(2), 115.

Scherthan, H., Weich, S., Schwegler, H., Heyting, C., Härle, M. and Cremer, T. (1996), 'Centromere and telomere movements during early meiotic prophase of mouse and man are associated with the onset of chromosome pairing.', *Journal of Cell Biology* **134**(5), 1109–1125.

Schwander, T., Libbrecht, R. and Keller, L. (2014), 'Supergenes and complex phenotypes', *Current Biology* **24**(7), R288–R294.

Shields, O. (1992), 'World distribution of the *Vanessa cardui* group nymphalidae', *Journal of the Lepidopterists' Society* **46**, 235–238.

Smit, A., Hubley, R. and Green, P. (2015), 'RepeatMasker Open-4.0'. **URL:** *http://www.repeatmasker.org*

Smith, G. R. and Nambiar, M. (2020), 'New solutions to old problems: Molecular mechanisms of meiotic crossover control', *Trends in genetics* **36**(5), 337–346.

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. and Saastamoinen, M. (2022), 'Improved chromosome-level genome assembly of the glanville fritillary butterfly (*Melitaea cinxia*) integrating pacific biosciences long reads and a high-density linkage map', *GigaScience* **11**(1), giab097.

Srygley, R. B. and Dudley, R. (2008), 'Optimal strategies for insects migrating in the flight boundary layer: mechanisms and consequences', *Integrative and Comparative Biology* **48**(1), 119–133.

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W. and Smadja, C. M. (2017), '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.

Talavera, G., Bataille, C., Benyamini, D., Gascoigne-Pees, M. and Vila, R. (2018), 'Round-trip across the Sahara: afrotropical painted lady butterflies recolonize the mediterranean in early spring', *Biology Letters* **14**(6), 20180274.

Talavera, G. and Vila, R. (2016), '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* **120** (2), 274–285.

Talla, V., Soler, L., Kawakami, T., Dincă, V., Vila, R., Friberg, M., Wiklund, C. and Backström, N. (2019), '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.

Talla, V., Suh, A., Kalsoom, F., Dincă, V., Vila, R., Friberg, M., Wiklund, C. and Backström, N. (2017), '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.

Tiley, G. P. and Burleigh, J. G. (2015), 'The relationship of recombination rate, genome structure, and patterns of molecular evolution across angiosperms', *BMC Evolutionary Biology* **15**(1), 194.

Tobler, A., Kapan, D., Flanagan, N. S., Gonzalez, C., Peterson, E., Jiggins, C. D., Johntson, J. S., Heckel, D. G. and McMillan, W. O. (2005), 'First-generation linkage map of the warningly colored butterfly *Heliconius erato*', *Heredity* **94**(4), 408–417.

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. and Wheat, C. W. (2021), 'A complex interplay between balancing selection and introgression maintains a genus-wide alternative life history strategy', p. 2021.05.20.445023.

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. and Papa, R. (2017), 'Complex modular architecture around a simple toolkit of wing pattern genes', *Nature Ecology & Evolution* **1**(3), 1–12.

Van't Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., Hall, N., Darby, A. C. and Saccheri, I. J. (2016), 'The industrial melanism mutation in british peppered moths is a transposable element', *Nature* **534**(7605), 102–105.

Wahlberg, N. and Rubinoff, D. (2011), '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.

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. and Paterson, A. H. (2012), 'MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity', *Nucleic Acids Research* **40**(7), e49.

Weber, J.-M. (2009), 'The physiology of long-distance migration: extending the limits of endurance metabolism', *Journal of Experimental Biology* **212**(5), 593–597.

Wells, J. N. and Feschotte, C. (2020), 'A field guide to eukaryotic transposable elements', *Annual Review of Genetics* **54**(1), 539–561.

Wickham, H. (2009), *ggplot2: elegant graphics for data analysis*, Use R!, Springer-Verlag, New York.

Wieczorek, H., Beyenbach, K. W., Huss, M. and Vitavska, O. (2009), 'Vacuolar-type proton pumps in insect epithelia', *The Journal of Experimental Biology* **212**(11), 1611–1619.

Wolfe, D., Dudek, S., Ritchie, M. D. and Pendergrass, S. A. (2013), 'Visualizing genomic information across chromosomes with PhenoGram', *BioData Mining* **6**(1), 18.

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. and Mita, K. (2008), 'A BAC-based integrated linkage map of the silkworm *Bombyx mori*', *Genome Biology* **9**(1), R21.

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. and Li, X. (2020), 'Chromosome-level reference genome assembly and gene editing of the dead-leaf butterfly *Kallima inachus*', *Molecular Ecology Resources* **20**(4), 1080–1092.

Yasukochi, Y. (1998), 'A dense genetic map of the silkworm, *Bombyx mori*, covering all chromosomes based on 1018 molecular markers', *Genetics* **150**(4), 1513–1525.

Zhan, S. and Reppert, S. M. (2013), 'MonarchBase: the monarch butterfly genome database', *Nucleic Acids Research* **41**(D1), D758–D763.

Zhan, S., Zhang, W., Niitepõld, K., Hsu, J., Haeger, J. F., Zalucki, M. P., Altizer, S., de Roode, J. C., Reppert, S. M. and Kronforst, M. R. (2014), 'The genetics of monarch butterfly migration and warning colouration', *Nature* **514**(7522), 317–321.

Zhang, L. (2003), 'Does recombination shape the distribution and evolution of tandemly arrayed genes (tags) in the *Arabidopsis thaliana* genome?', *Genome Research* **13**(12), 2533–2540.

Zhang, L., Steward, R. A., Wheat, C. W. and Reed, R. D. (2021), 'High-quality genome assembly and comprehensive transcriptome of the painted lady butterfly *Vanessa cardui*', *Genome Biology and Evolution* **13**(7).

Zhu, H., Gegear, R. J., Casselman, A., Kanginakudru, S. and Reppert, S. M. (2009), 'Defining behavioral and molecular differences between summer and migratory monarch butterflies', *BMC Biology* **7**(1), 14.