Table of contents

Genome Biology

https://genomebiology.biomedcentral.com/submission-guidelines

Abstract

(250 words)

- Background: the context and purpose of the study
- Results: the main findings
- Conclusions: a brief summary and potential implications

Keywords

Background

Era of genomics opens up opportunities for studying relationships between genotypes and complex phenotypes on a novel level. Arising sequencing technologies and their combinations can guide the researchers in understanding genome evolution and therefore genomic patterns specifically characterizing complex phenotypes. We now are able to follow accumulation of change through the process of adaptation on different levels: gene duplications, accumulation of mutations, process of recombination, evolution of selfish elements (transposable elements). The key to understanding the above mentioned processes is high quality complete genome assembly. Special level of resolution can be achieved when we are able to follow the history of the chromosomes as units, therefore requiring chromosome level assembly. One of the most powerful methods to achieve both chromosome level assembly and ensure it's spatial correctness is linkage map. In this study we present the linked chromosome

level assembly of the Painted lady *(Vanessa cardui)* genome - species extraordinary in many aspects: long-distance migration, high population size etc. We provide first insights on how various genetic mechanisms contribute to the evolution of the genome and influence formation of the complex phenotype. Below we discuss in more detail how different mechanisms influence evolution and introduce the study system.

Gene duplication have long been...

Recombination: how? Recombination is a process of great significance in evolutionary biology. This influence rises from the capacity recombination has of generating novel haplotypes in the offspring, as well as breaking down previously existing adaptive allele combinations. Such conflict between the possible outcomes is fundamental for many core questions, e. g. level of genetic diversity, the evolution of sexual reproduction and the establishment of reproductive barriers that drive speciation. (AP) Recombination rate is negatively correlated with chromosome size, expected if only 1-2 rec events per chromosome (Figure C, Table C). Also observed in Heliconius (ref)

Transposable elements (TEs) are mobile DNA sequences capable of independently replicating within host genomes. They typically range in length from 100 to 10,000 bp, but are sometimes far larger (6). Along with viruses, TEs are the most intricate selfish genetic elements. doi.org/10.1146/annurev-genet-040620-022145

Interplay of all the above mentioned mechanisms is important and poorly understood. In addition, genomic regions experience various strengths of pressure from different evolutionary mechanisms. One of the charismatic examples is sex chromosome evolution, where all the above mechanisms may function in a specific way.

Broad view and combination of different approaches is of high importance for studying evolution of the comprehensive phenotypes. The Painted Lady butterfly (*Vanessa cardui*) is one of the species of high interest for evolutionary studies, due to multiple outstanding traits. This is the most cosmopolitan of all butterfly species (Talavera et al., 2018) and it's migratory behaviour presents a diverse repertoire of phenotypes. In it's long multigenerational journey it faces not only pressures of long-distance flight, but also extreme environmental heterogeneity. Yet Vanessa cardui appears to successfully adapt to all the selective pressures imposed by migratory lifestyle and maintain large population size. Until now we had very few insights how this is manifests of the level of the genome.

In this study we achieve three goals: contribute to community effort to bring the field of evolutionary biology to the era of genomics by providing linkage map correction of the Vanessa cardui genome. Among insects, genome assemblies are currently available for 401 species (Li., 2019), however their quality varies significantly and annotations are available for just around 13% for the species (more on Lepidoptera genomes in Triant et al, 2018).
Secondly, we lay out foundations to study complex migratory phenotype from various angles: evolutionary genomics, epigenetics and population genetics. We investigate evolution of gene families and discuss it's adaptive characteristics. Access to linkage map provides opportunity for

deeper analysis of recombination rate variation and high resolution of the recombination map is crucial for population genetics approach. Epigenetics relays on high quality and high contiguity of the genome and gene annotation. Finally, our investigation of evolutionary patterns brings new insights on evolution of butterfly genomes (Lepidoptera) in general. Questions like … have never been addressed for this entire class.

Results

Basic genome statistics

We verified the existing assembly from the Darwin Tree of Life project with a linkage map. Genome occupies total length of 424Mb and consists of 30 autosomes and both sex chromosomes Z and W. Large N50 and high BUSCO scores allow us to suggest that chromosomes are nearly complete.

Total length of sequences marked as repeats slightly exceeded 100Mbp (23.38%). Results are summarized in the Suppl. table.

Genome	Total sequence length	424 Mb
	Number of chromosomes	32
	Scaffold N50	14Mb
	BUSCO genes	97%
	GC content	
Repeats	Percent of repeats	23.4%
Genes	High quality gene models	14560
	Genes with functional annotation	12000

Synteny

We accessed collinearity of the *Vanessa cardui* genome using predicted gene models and synteny alignment using MCScanX. Collinearity of the genome is tested on two levels of genetic divergence: we first compared V.cardui genome to the genome of Silk Moth *Bombyx mori*. On

the closer evolutionary scale we performed a comparison with other species from Nymphalidae family: postman butterfly (*Heliconius melpomene).*

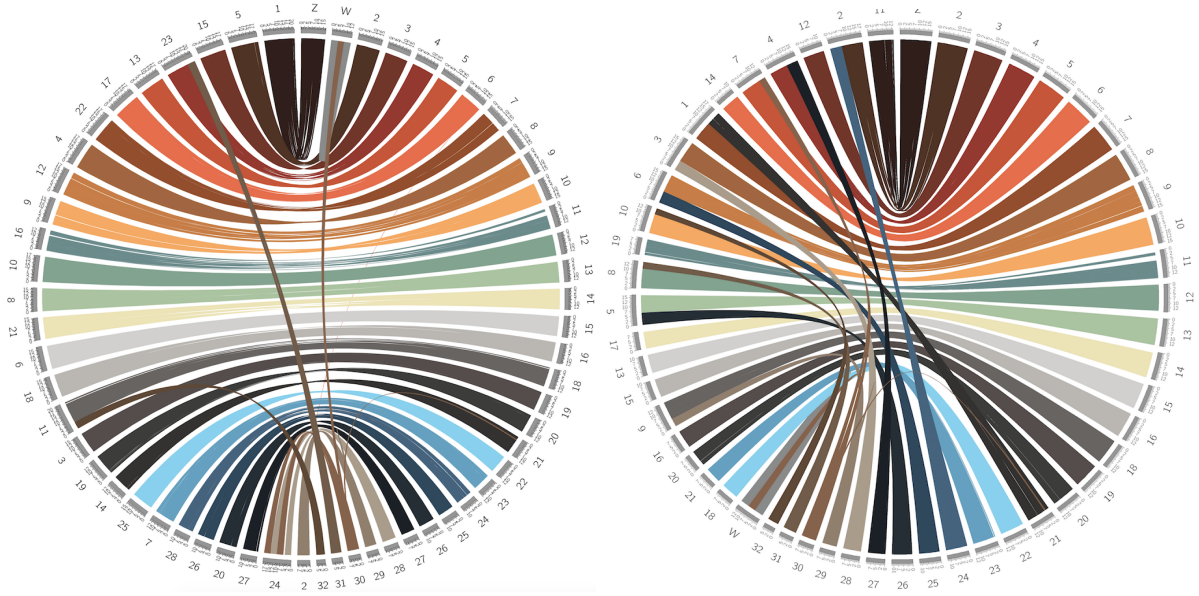


Figure H. Visualization of the synteny alignments between *Vanessa cardui* and a) *Bombyx mori*, b) *Heliconius melpomene*

Overall we observe high levels of synteny between all the compared species. Notable exception from this pattern is W chromosome, which is not present in the genome of neither *B.mori* nor *H.melpomene.* Chromosome 30 of *V.cardui* has syntenic regions in multiple chromosomes and may indicate scaffolding error or complexity of this region, which confounds genome assembly. In agreement with the previous studies we observe fusion of the chromosomes in the *H.melpomene* genome.

Gene family evolution

OrthoFinder clustered 93.8% (159472) of the genes from the nine Nymphalid species in 15295 orthogroups. The percentage of genes assigned to orthogroups varied from 89.8 to 99.6% for different species. In *V. cardui* 95.6% (13913) of the genes were assigned to 10782 orthogroups with 27 species-specific orthogroups containing 128 genes (Table SX). The composite species tree was in accordance with earlier published species trees from the literature (Espeland et al. 2018). This is of importance when inferring duplication events on the different branches. OrthoFinder inferred 86 duplications on the common *Vanessa* branch and 1125 duplications on the tip branch to *V. cardui* compared to 6218 in *V. tameamea* (Figure A).

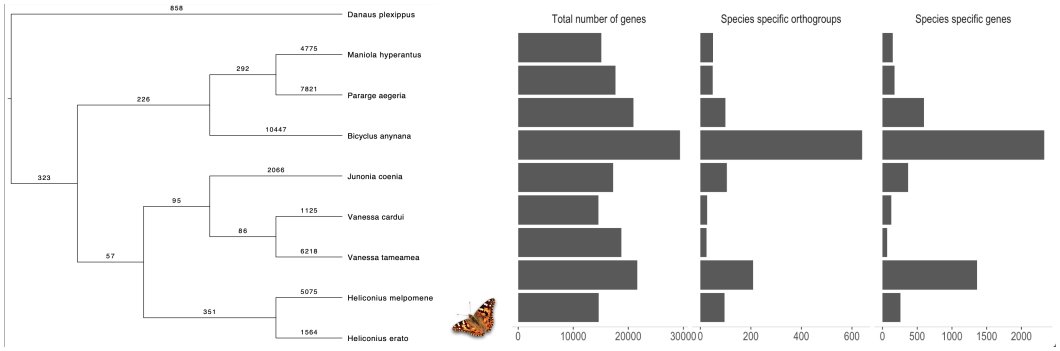


Figure A. Gene family evolution on Nymphalid species tree inferred by OrthoFinder. The number on the branches shows duplications. The barplots show the total number of genes included, species-specific orthogroups and number of genes in species-specific orthogroups.

The maximum likelihood rate estimation resulted in 292 orthogroups having different gene family expansion/contraction rate in the *V. cardui* compared to the other branches. Of these, 83 orthogroups containing 758 genes experienced gene gains and 25 orthogroups had experienced gene losses (Figure B). Significantly enriched GO-terms associated to genes undergoing expansions in *V. cardui* are displayed in figure B, noteworthy is that 11 of the 22 significantly enriched GO-terms in the category biological functions involved fatty acid metabolism and three involved regulation of gene expression (Figure B).

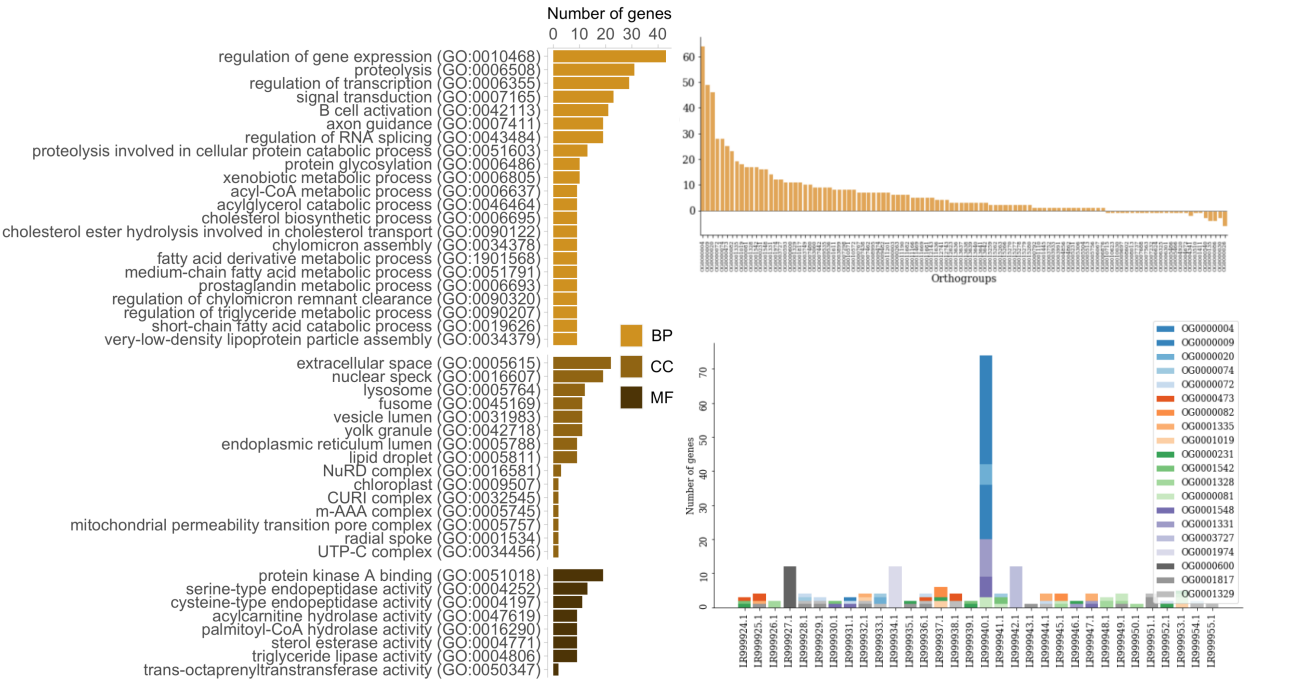


Figure B. A) Enriched gene ontology (GO) terms associated to the genes gained in the V. cardui branch,

with p-value < 0.05 after FDR-correction. The bars show the number of genes associated to each GO-term. The different GO-categories are biological processes (BP), cellular compartment (CC) and molecular function (MF). B) Orthogroups that experienced gene gain or loss according a model with distinct rate of gene expansion and contraction in the *V. cardui*-branch. C) Chromosomal distribution of the expanding gene families (20 most common orthogroups) with number of genes on the y-axis and the colour displaying orthogroup affiliation.

Testing a branch model where orthogroups have been experiencing the same rate in *V. cardui* and *D. plexippus* but not in the other nymphalids resulted in 39 gene families expanding in both branches. These orthogroups included 196 genes, with functional enrichment of various metabolic processes and multiple terms involved in neurotransmitter activity (Figure SX). We also investigated the presence of specific expansions for the branch leading to the *Vanessa*-genus. There were 49 orthogroups that showed expansions distinct to the *Vanessa*-branch comprising 246 genes. These were functionally enriched in pathways associated with heart rate and voltage-gated ion channel activity, other functions included regulation of acetylcholine activity, ovulatory cycle rhythm and juvenile hormone regulation.

Location of expanded families along the genome.

Identification of candidate genes of interest in the monarch.

Patterns of recombination

The total map length is 1375 cM, and the maps were constructed with 1696 markers with average marker density of 3.95 markers/Mb. The average recombination rate calculated as the map length/genome size resulted in an average genome wide recombination rate of 3.19 cM/Mb. Excluding the W_chromosome the global recombination rate was 3.42 (sd +/- 0.72) and the mean recombination rate per chromosome varied between 1.75-5.34 cM/Mb. The recombination rate in the Z-chromosome was 2.48 cM/Mb, which is less than the average recombination rate among the autosomes (unweighted), but not below the expected recombination rate for its size based on the regression line.

Correlation to chromosome length: Recombination rate is negative correlation with chromosome length, this is also seen in overall repeat abundance, repeat proportion and GC-content. Gene number is positively correlated to chromosome length.

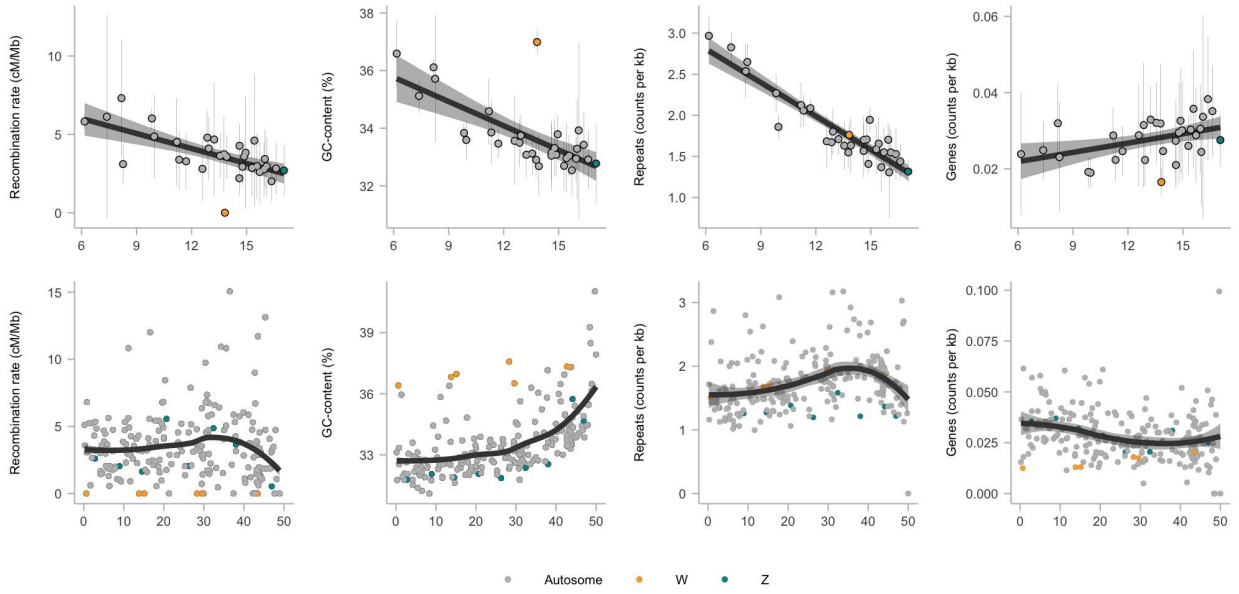


Figure C and D. Correlation between chromosome length and genomic features. Regional distribution

The regional recombination rate was estimated with non-overlapping window-based linear regression, with window size 2 Mb. The global recombination rate average across all windows in the genome was estimated slightly higher with this method 3.5 cM/Mb (sd +/-2.52), with a range of 0-15.5 cM/Mb. The chromosome average recombination rate ranged between 1.84 - 7.13 cM/Mb. The recombination rate was not uniform across the chromosome, there was a regional variation with lower rec rate in the centre of the chromosome and increasing in the flanking regions and then again reduced at the end of the chromosomes. Binning of the markers in five distance intervals from the centre of the chromosome resulted in a significant difference in mean rec rate in the terminal flanking intervals compared to the centre and terminal regions (stats). But there was no significant difference between the centre and the terminal regions.

The regional distribution of repeats follow the same pattern as the recombination rate, this is also observed when looking at the different repeat classes except in the LTR-elements and the TcM-element. We correlated the binned average recombination rate in 2 Mb windows across the genome to repeat content, gene content and GC-content. There was a significant positive correlation between recombination rate and total repeat abundance across the genome (Spearman's rank corr rho 0.28, p-value). The proportion of repeats showed a weak positive correlation with rec rate (rho 0.13, p-value), if excluding W the correlation became stronger (rho 0.25, p-value). Looking at different repeat classes the SINE and DNA-repeats are moderately positively correlated to recombination rate, none of the other classes show a significant correlation, but when excluding W all repeat classes have a significant positive correlation with

recombination rate (suppl figure, table). A moderate negative correlation between length per repeat and recombination rate disappeared when excluding W (suppl table).

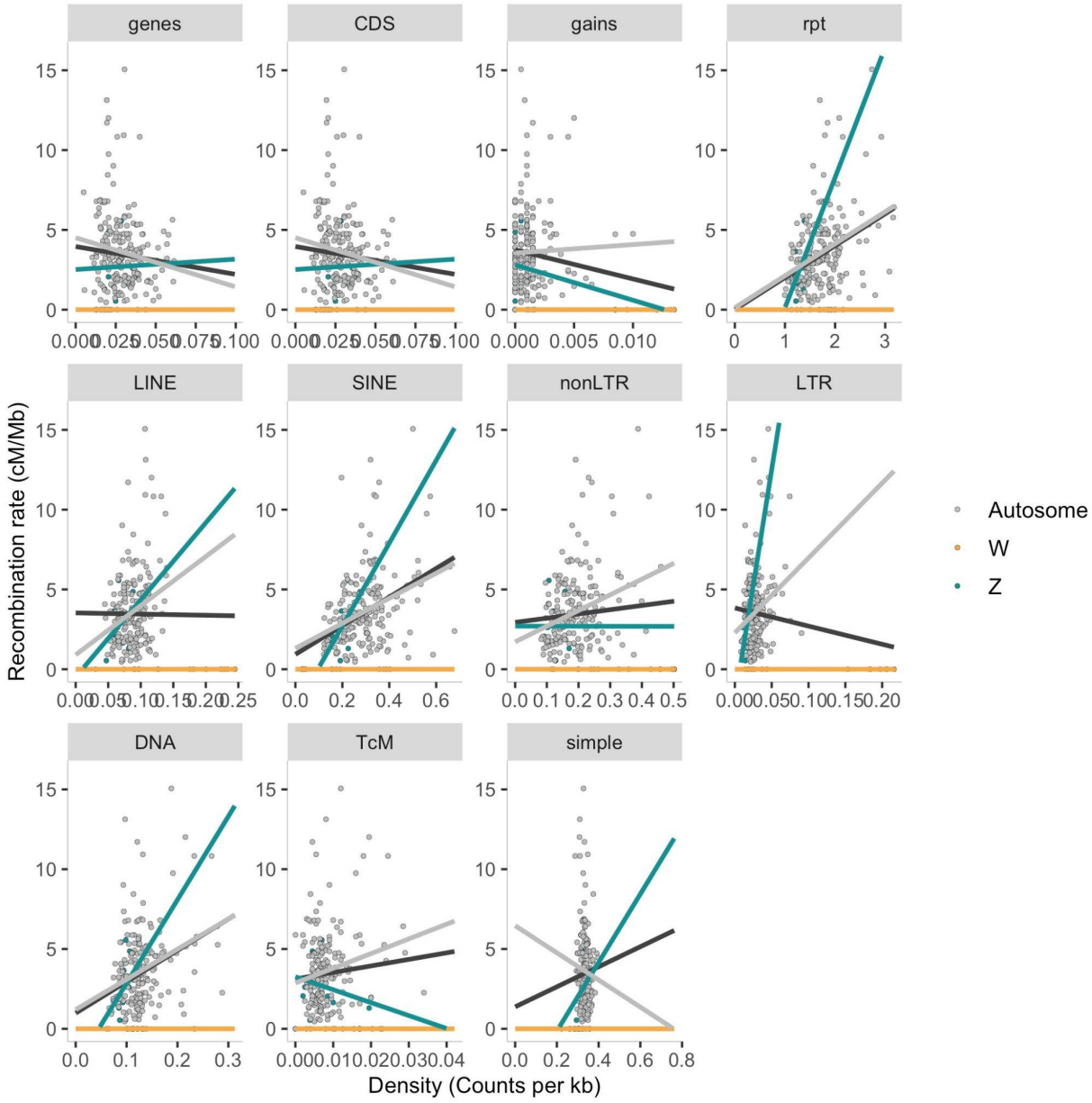
Genes and GC-content present a different pattern where the values increase at the end of the chromosomes. The gene abundance and proportion as well as the GC-content showed no significant correlation to recombination rate. The gene density is negatively correlated to GC-content (-0.21, p-value = 0.001519), but the proportion of genes is positively correlated to GC-content (0.14, p-value = 0.02971).

Using recombination rate as dependent variable and chromosome length and type, relative binned chromosome position, GC-content, gene density, density of gained genes and density for different repeat classes as explanatory variables, showed significant effects of the 4th position but not the length of the chromosome (est 1.6019±0.64328, t-value 2.490, p-value 0.0136). The SINE:s have a positive effect (lm estimate 1.07328±0.50957, t-value 2.106, p-value 0.0364, lmer est 1.248280±0.456803, t-value 2.733) and the density of non-LTR have negative effect (lm estimate -1.68622±0.78924, t-value -2.137, p-value 0.0339, lmer estimate -1.942449±0.694761, t-value -2.796) (lm F-statistic: 4.213 on 16 and 200 DF, p-value: 5.112e-07, R2 0.2521, adj R2: 0.1922) (suppl table).

The relationship between gene gains and genomic features was explored on in 2Mb window size to be able to include the rec rate as explanatory factor, chromosome type was tested as interaction with the other variables (R2: 0.7037, AdjR2: 0.654, F-statistic: 14.17 on 31 and 185 DF, p-value: < 2.2e-16). The recombination rate does not appear to have any association to gained genes. The density of genes is just non-significant in the autosomes with this window size, and has no significant effect on the gene gain distribution on the W or Z-chromosomes. The factors significantly associated to gene gains are the density of LINE:s (est 0.383979±0.122578, t-value 3.133, p-value 0.00202), the density of SINE:s (est -0.452578±0.143681, t-value -3.150, p-value 0.00191) and density of DNA-elements (est 0.333417±0.108672, t-value 3.068, p-value 0.00248). When comparing the mean of the counts in each window and comparing those with and without gene gains, the features with significant difference were LINE:s and LTR, and when excluding W only LTR retained a significant difference in means (suppl table).

We then compared the features on a detailed level in 100 kb windows within a linear model, and now there is a small effect of gene density on the number of genes gained on the autosomes. There was no significant effect on gene gains on the Z-chromosome. Other significant effects on the number of gained genes on the autosomes are a negative effect of SINEs and a positive effect of LTR. DNA-element, non-LTR and LINEs have an effect on gained genes on both

autosomes and W-chromosome. Simple repeats are associated with gene gains on the W. (R2: 0.1518, AdjR2: 0.1466, F-statistic: 29.15 on 26 and 4235 DF, p-value: < 2.2e-16) (suppl table).



E Suppl fig. Corr rec rate vs genomic features

Chromosome level assembly enables us to look at Gene duplications/gains throughout the genome. We observe three patterns: orthogroups spread between several chromosomes (how many?), orthogroups in clusters or within the same chromosome. (Also Figure F).
Results of stats genegain vs no gains

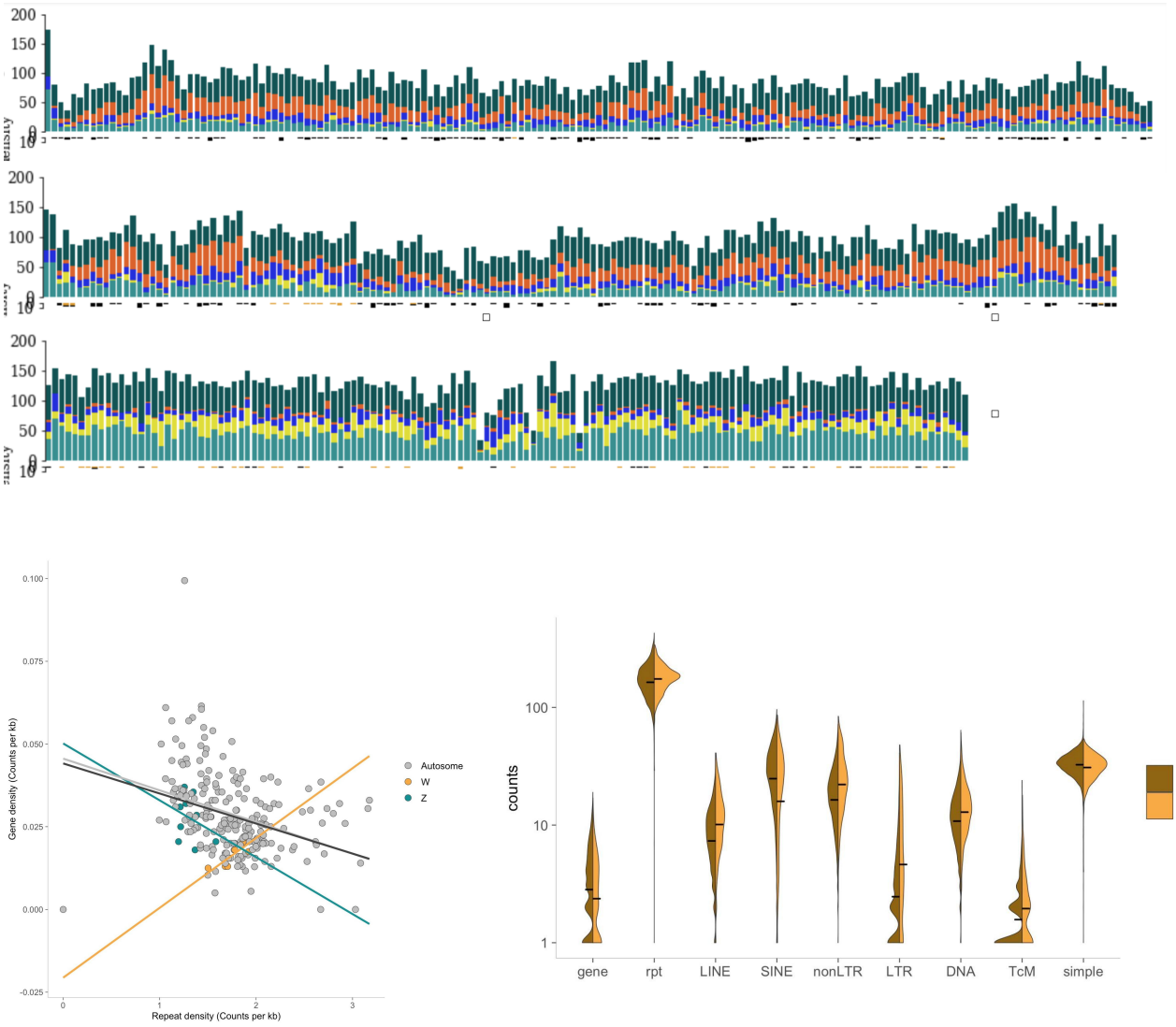


Figure F (preliminary). Window-based analysis: distribution of repeat classes, genes and genes belonging to expanded gene families along the chromosomes. Figure G. Correlation between genes and repeats. Figure H. Violin plot of gene gains vs no gene gains in different genomic features.

Discussion

Summary of this chapter

Vanessa genome is shaped by various evolutionary forces and processes.

1. Basic genomics and synteny analysis confirms high synteny of Vanessa cardui genome
2. Adaptation to migratory lifestyle in two forms
    a. Adaptation to migration itself (fat metabolism)

        i. It's unlikely that adaptation of highly migratory lifestyle involves the same set of genes in different species (Monarch butterfly)
    b. Adaptation to wide range of host plants (we see signatures of it in expansion of gene regulation genes)
    c. Vanessa is evolved to adapt and be plastic
3. Recombination landscape in Vanessa cardui
4. Evolution of sex determination system
5. Gene duplication promoted by recombination and/or TE

Basic genomics and synteny analysis

a. Major challenges in insect genome assembly projects are dealing with repetitive regions and high levels of heterozygosity. We overcome those by using a combined approach: long-read sequencing, scaffolding with Hi-C and linkage map.

1. Construction of linkage map is of particular importance for this paper, since it Vanessa genome is shaped by various evolutionary forces and processes: a) verified chromosome level assembly, b) gives new insights to pattern of recombination, c) opened up possibility for all consecutive analysis of correlations between genomics features, d) allowed reliable analysis of spatial distribution
    a. Methodological advances?
2. On the other hand we assume specifics of the Vanessa: migration, high adaptability, high Ne, variety of host plants.
3. We first determine place of the Vanessa in the butterfly genomics:
    a. Vanessa has average genome size and number of chromosomes corresponds to ancestral state (Table 1)
    b. Repeat content is relatively low, comparing to Lepidoptera (Table 1)
    c. Number of genes is slightly smaller than core gene set (Table 1)
    d. Vanessa cardui genome demonstrate high synteny, like the other Nymphalid genomes (Fig. 1 Synteny).
    e. W chromosome is not present in majority of available genomes and inferring it's origin and synteny. However, we observe very little connection with other known chromosomes from Bombyx and Heliconius.
4. Number of duplications on Nymphalid tree is very variable and Vanessa cardui is on the lower end of the spectrum (Figure A). Consequence of high Ne! 5000000 Ne.mutation rate?

Adaptation to migration

1. We use analysis of gene gains as a way to determine orthogroups potentially evolved in response to adaptation. Number of gene families gained and expanded in the Vanessa cardui genome are involved in fat metabolism. It was previously shown that accumulation of fat is of particular importance for long range migration. (Figure B a)

The Painted Lady (*Vanessa cardui*) is the most cosmopolitan of all butterfly species and migratory behaviour presents a diverse repertoire of phenotypes. It outnumbers the migrations of some well-known mammals and birds (Holland et al. 2006). Vanessa cardui inhabits an impressive range of habitats from temperate to tropical, making it an unprecedented case for butterflies and most of the insect species (Talavera et al., 2018). Developmental time is also highly dependent on local conditions, mainly temperature (Stefanescu et al., 2013). This striking environmental heterogeneity can, in principle, impose different selection pressures on the migrating population at different points of the annual journey.

2. Genomics of migration is investigated in great detail for Monarch butterflies. One can assume that some of the genetic mechanisms can be reused in Vanessa. However, it's unlikely that adaptation of a highly migratory lifestyle involves the same set of genes in different species (Monarch butterfly). We show it by comparing migratory gene dataset with Vanessa cardui gained genes.

Evolution of high adaptability

1. We see signatures of high adaptive capability in expansion of gene regulation genes (Figure A a)
2. This could be a signature of capability to adapt to: host plants, various habitats, different environmental conditions, temperature. Local adaptation is useless, because environments are drastically changing between generations.
3. We don't use substitutions

Processes of recombination

1. Recombination rate is there (its vital for all the evolutionary research), we went beyond that and investigated factor which may shape recombination landscape in *V.cardui.*
    a. We are one of the first estimating rec rate in Lepidoptera. Rec rate is slightly lower in Vanessa compared to Heliconius. However, it's difficult to compare recombination maps made with different methods. Number and size of chromosomes differ between Helic.and Vanessa.

    As for the absolute value of recombination rate, the genome-wide average in *L. sinapis* seems to be around 5 – 7 cM/Mb. This is substantially higher than the usual vertebrate rates (Beye *et al*. 2004; Supplementary Table 1 and references therein), ranging from 0.16 cM/Mb in the Atlantic trout to 3.17 cM/Mb in chicken, with humans presenting an intermediate recombination rate of 1.22 cM/Mb. It is in line with previous measurements in insect species from different lineages, higher than in Diptera including both mosquitoes and *D. melanogaster*, and lower than in the honey bee, which presents an exceptionally high genome-wide recombination rate at 19 cM/Mb (Beye *et al*. 2004). When comparing with other Lepidoptera, the genome-wide recombination rate in the wood white butterflies is on the same level. *B. mori* has an average recombination

rate of 4.6 cM/Mb (Yasucochi 1998) while *Heliconius erato* shows 6 cM/Mb (Tobler *et al.* 2005) and *H. melpomene* 5.5 cM/Mb (Jiggins *et al.* 2005). (AP)

b. Variation in recombination rate is suggested to be a trait under selection (ref). A high recombination rate could potentially increase diversity and efficiency of selection (ref). However, a species with very high heterozygosity (and a high estimation of Ne) the selection pressure to evolve a high recombination rate is low. There is possibly even a reversed selection pressure to reduce the number of recombinations since recombination could have negative effects, such as uncoupling of beneficial allele combinations and mutagenicity (ref).
    i. Unfortunately, we don't have population data to test this hypothesis.

2. Earlier studies have shown a correlation between recombination rate and chromosome size, we observed this but when looking across the whole genome the effect of chromosome size is not significant. The only predictors significantly associated with recombination rate are the position on the chromosome, density of SINE:s and non-LTR:s.

   A classic example involves several families of LINEs (e.g., R1, R2, etc.), which precisely target ribosomal RNA (rRNA) gene arrays (34).

   If this model is correct, then longer TEs should be strongly selected against due to their increased likelihood of initiating recombination. Indeed, LTR and LINE retroelements tend to fix and cluster in regions with low recombination rates (e.g., peri-centromeric heterochromatin), while shorter elements such as SINEs and MITEs accumulate in gene-rich regions, which are generally characterized by higher recombination rates (22, 33, 165). The relationship between TEs and recombination is a complex one, however, and is discussed in more depth elsewhere (81).

   a. We first observed a negative correlation between length of the chromosomes and recombination, also observed in Heliconius (ref).
   b. Overall, we observe weak positive correlation between gene density and chr size. In birds neg corr gene density and chr size? (Figure C). No correlation to coding sequence and recombination rate.
      i. The hypothesis is that recombination rate is positively correlated with gene density.
      ii. For our data, there is no significant correlation between recombination rate and gene density or CDS content. (Figure E)
      iii. However, we observe the opposite spatial pattern in gene and GC-content distribution throughout the chromosome: it increased towards ends (Figure D)
   c. One of the striking results is that we have strong neg correlation with repeats and chr size too. (Haven't been addressed before?)

3. Linear model results

a. Recombination rate unevenly distributed along the chromosome, lower in the centre, higher in the flanking regions and lowers towards the chromosome ends (Figure D). In contrast with others (ref).
    i. Consistent with the severely deleterious consequences of improper chromosome segregation, recombination rate drastically reduces in the pericentric and most distal regions in the chromosomes (Smith & Nambiar 2020). In holocentric chromosomes we would expect even distribution due to mechanical reasons. But it might not explain lower values of rec rate in the center of the chromosome.
    ii. The distribution of recombination events along the different scaffolds seems to be quite different than what would be expected from monocentric organisms (species whose chromosomes have a single defined centromere). In these species, recombination is completely prevented in the centromere, while the rate progressively increases toward the telomeres. This has been described, for example, in insects such as *D. melanogaster* (Hey & Kliman 2002) and honey bee (Beye *et al*. 2006), and mammals like mouse, rat and humans (Jensen-Seaman *et al*. 2004). In the case of holocentric species, where the centromere effectively takes over the whole length of the chromosome, this relationship is, at the very least, more difficult to establish. However, even in *C. elegans*, with holocentric chromosomes, a similar pattern has been obtained, with recombination happening more often in distal positions compared to more central ones (Prachumwat *et al*. 2004). (AP)
    iii. Interference might explain lower rec rate in the center of the chromosome?
    iv.
b. Positive correlation between recombination rate and SINE and (neg) with other non-LTRs. (Figure E, Linear model)
    i. non-LTRs: Prevailing hypothesis in the literature is that recombination rate is negatively correlated with recombination rate (ref). Recombination facilitates the removal of repeats which are generally deleterious.
    ii. Discuss ALU (SINE inducing recombination in humans)
    iii. Our results are in contrast with this hypothesis: here is a positive correlation between all repeat classes and recombination rate (excluding W). (Figure E)

All the above are likely to jointly influence evolution of Vanessa genome

(Global picture, including all the features)

1. (Figure B c) Chromosome level assembly enables us to look at Gene duplications/gains throughout the genome. We observe three patterns: orthogroups spread between

several chromosomes (how many?), orthogroups in clusters or within the same chromosome. (Also Figure F)

2. Is pattern caused by the mechanism of duplications? “Chromosomes evolve by transposition of mobile elements; by gross rearrangements such as inversions, translocations, deletions, and duplications; by homologous recombination; and by slippage of DNA polymerases during replication. It is likely that all of these mechanisms have contributed to the proliferation and dispersal of protein building blocks”
    1. Our findings support the first scenario with mobile elements.
        a. SINEs are negatively correlated with gene gains
        b. LTR, non-LTR, LINE, DNA positive
        c.
    2. Recombination and gene gains
    3. Spatial distribution of repeats (Figure H) demonstrates universal patterns throughout the genome with and exception of W chromosome

    Further study: Timing of the gene duplications (divergence)
    1. W chromosome harbors significantly more gene duplications, due to …

1. Z chromosome is the largest chromosome, no significant difference in gene gains.
2. Recombination rate in the Z-chromosome is lower than the average autosome rec rate (Figure C, Table C). Not expected based on no recombination in females (Z ⅔ in males rec rate should be higher).The Z-chromosome is the largest chromosome that could counteract the higher expected recombination rate in the Z-chromosome.
3. W chromosome demonstrates a number of highly specific features: different number of repeat classes compared to other chromosomes (Figure Suppl barplot), gained genes.
4. We think the chromosome may be young, because:
    a. It’s long
    b. Harbors repeats and genes (sometimes with unknown function).
    c. Repeats are longer on W (Suppl. Fig)

Conclusions

6. Chromosome level assembly verified by linkage map serve as a resource for genome architecture, organization, population level studies etc
7. Vanessa genome evolved under selection to adapt to migratory lifestyle and due to intreply between different forced effecting it: recombintion, repeats
8. Lots of fat and regulatory genes
9. Recombination landscape in Vanessa cardui

10. W chromosome is very special and opens up opportunities to study chromosome evolution on the large scale (Lepidoptera)
11. Gene duplication promoted by recombination and/or TE?

To conclude: we provide new version of the genome which is the great tool and also new insights on lepidoptera genomics as a whole and deep analysis of genome evolution of intriguing system (second after the monarch)

Methods

Linkage map

Sampling and DNA-extraction

Offspring (> 100) from one female caught in El Brull in Catalonia were reared on cuttings of thistles in the greenhouse until pupation. To confirm that only one father had sired the offspring, the genitals of the female were examined and only one spermatophore detected. The offspring were snap frozen in liquid nitrogen and stored in -20°C until extraction. DNA extraction with a modified high salt extraction method was performed on the thorax of the female and on the upper abdominal segment of the offspring (Aljanabi 1997) . The amount and quality of the DNA was analysed with Nanodrop (ThermoFischer Scientific) and Qubit (ThermoFischer Scientific). The DNA was digested with EcoR1 enzyme according to the manufacturer's protocol with 16 hours digestion time (ThermoFischer Scientific). The efficiency of the digestion was determined by visual inspection of the fragmentation on gel electrophoresis. We selected high quality digested DNA from 95 offspring together with the female parent for RAD-sequencing of 2x151 bp paired reads on one lane with NovaSeq6000 at the National Genomics Infrastructure, SciLife, Stockholm.

Data processing

We predicted the expected number of EcoR1 enzyme cut sites in the genome by using PredRAD (Herrera et al. 2015). The quality of the raw reads was initially assessed with FastQC (Babraham Bioinformatics and Andrews,S. 2010. FastQC A quality control tool for high throughput sequence data. April 26, 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). We filtered the reads using the Stacks2 modules clone_filter to remove PCR-duplicates and process_radtags to remove reads with average phredscore <10 (90% probability of correct base called) in windows 15% of the length of the read (Catchen et al. 2013). Additional filtering was applied by using options -c to remove all

reads with unassigned bases and truncating the reads to 125 bp. The option --disable_rad_chec was applied to keep reads without complete radtags.

We mapped the filtered reads to the genome produced by the Darwin Tree of Life initiative (available in the NCBI database, genome GCA_905220365.1_ilVanCard2.1_genomic.fna.gz, accessed 13/03/2021) using bwa mem algorithm (Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN].) with default options and quality filter –T (only output reads with mapping quality score above 30), and sorted resulting bam files with samtools sort (http://bio-bwa.sourceforge.net, Li et al. 2009). The bam files were further filtered with samtools view –q 30 option and a custom script removing all reads with flags indicating multiple mappings "XA:Z:" and supplementary hits "SA:Z:" so that only reads with unique hits were retained. The mapping coverage was analysed with Qualimap (Okonechnikov et al. 2015). The offspring was 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 with parameters -q 10 -Q 10 (Li 2011). The variants were then converted to likelihoods with Pileup2Likelihoods in LepMap3 using default settings of 3 reads as minimum coverage per individual, 30% individuals allowed with lower coverage, and a minimum allele frequency of 0.1 (Rastas 2017).

Construction of linkage map

LepMap3 was used to construct the linkage map (Rastas 2017). The module ParentCall calls informative parental markers and uses genotype likelihood information from the offspring to impute missing or erroneous parental markers. This was run with default values, except for zLimit=2 to detect markers segregating as sex chromosomes, and set to remove non informative markers, resulting in 7921 (42849) markers. The markers were assigned to linkage groups using SeparateChromosomes2 with informativeMask=2, lodDifference=2 and distortionLod=1, only using maternal informative markers to create robust groups. The LOD-limit was empirically estimated by testing a range of LODscores (1-30) and set to 12 resulting in 31 linkage groups.

JoinSingles was run without informative mask, ie using all markers, and lodLimits 10 deemed optimal, to assign all informative markers to linkage groups resulting in 6188 (23076) markers assigned to linkage groups. OrderMarkers was run with 50 iterations for each linkage group to determine the most likely distance between the markers in a maximum likelihood framework and the maps with the highest likelihood was selected for refinement. Only male informative markers (informativeMask=1) were included in the ordering of the map, since female informative markers do not contain any information on recombination events and only add noise. Additional options used were the Kosambi distance method, minError=0.1 and recombination2=0. The 30% end of the linkage groups was assessed and any marker or group of markers more than 10cM from the nearest marker was trimmed. The trimmed map was then reevaluated with MarkerOrder with the same settings. Uninformative markers at the map ends were manually removed. The maps were

then reordered again by OrderMarker evaluate order function to obtain the final map distances. The maps were thinned so that only one SNP per 200 bp were left (i.e. at least one SNP per radtag). Any remaining uninformative markers at the map ends were manually removed after visual inspection. To anchor the markers to the preliminary assembly we used the software lepAnchor (Rastas 2020). The workflow wrapper lepanchor_wrapper.sh was used with default settings.

Genome annotation and whole genome statistics

Genome assembly statistics

Anchoring of the linkage map markers to the *Vanessa cardui* genome version from the Darwin Tree of Life (DToL) initiative showed perfect alignment and confirmed structure of all fully assembled chromosomes. For the further analysis we didn't include unassembled haplotigs from DToL version of the genome assembly. We calculated summary statistics of the genome after linkage map verification using QUAST suite (10.1093/bioinformatics/btt086) with BUSCO (https://doi.org/10.1093/molbev/msab199) gene analysis option.

We used MCScanX (ref) software to describe syntenic blocks between *Vanessa cardui* genome and *Bombyx mori* and *Heliconius.* BLAST was used for preliminary alignment to serve as an input for the software. We used circos library (ref) for visualization of results.

Gene and repeat annotation

The annotation of the V.cardui genome was performed with the MAKER package, version 3.00.0 (https://doi.org/10.1186/1471-2105-12-491). We executed the MAKER pipeline iteratively in three stages. At the first step we masked repeated sequences and mapped transcriptomic evidence to the genome version verified by linkage map. RepeatMasker version 4.0.3 (Smit et al., 2013-2015) was used within the MAKER pipeline with manually curated Lepidoptera repeat database (doi.org/10.1093/gbe/evx163) serving as a reference. Additionally, RepeatMasker produced annotation of the repeats, including their position along the genome and classification into classes. Resulting file was further used for window-based and correlation analysis.

At the first MAKER run we used transcriptomic data from V.cardui wing transcriptome (https://doi.org/10.1186/s12864-016-2586-5) accessed on (). This step produced set of gene models, which we controlled for quality 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 corresponding evidence sequence) using custom scripts. Resulting gene models provided as a training set for the second run of MAKER.

The second iteration of MAKER pipeline was used to create gene models using the ab-initio gene predicting algorithm implemented in SNAP (https://doi.org/10.1186/1471-2105-5-59).
For the last run of MAKER we used gene models predicted by SNAP and additional protein evidence from Uniprot database (https://www.uniprot.org/; accessed 2021-04-01). We downloaded a set of Lepidoptera proteins from the Swiss-prot section of the Uniprot database and curated it manually. All the genes from the "reviewed" set were included, from the "unreviewed" set we selected only fully sequenced nuclear proteins with predicted functions (custom scripts were used for selection). This selection resulted in 36,907 proteins.
Finally, all obtained evidence and predicted genes were merged resulting in 18,860 gene models. Analogously to the first step, we set AED score to 0.5 and filtering reduced number of gene models to . Resulting genes were renamed using MAKER supplementary scripts.

Functional annotation of V.cardui was performed using eggnogg_mapper online tool (Huerta-Cepas,). Eggnogg assigns functional information to the genes using orthology information available in an integrated precomputed database. When orthologs are identified eggnogg assigns functional information from GO, Pfam and KEGG databases. We recovered functional information of 14500(?) genes. We controlled quality of orthologs alignment using custom filtering and the resulting dataset consisted of 12000(?) genes.

Gene family evolution

We investigated gene family evolution in *V. cardui* by comparing our newly obtained gene annotation with other annotated Nymphalid genomes available on Lepbase. The protein fasta files were downloaded 210621 (http://download.lepbase.org/v4/sequence/) (Table SX, versions of all used genomes, incl result from Orthofinder). To cluster the annotated genes into orthogroups and infer species specific orthogroups and gene duplications we used OrthoFinder/2.5.2 with default settings (Emms and Kelly 2019). The total gene counts for each orthogroup and species from OrthoFinder was used as input to estimate gene family expansion and contraction with the software Badirate using a maximum likelihood option and the birth/death/innovation (BDI) model (Librado et al. 2012).

Badirate requires an ultrametric tree as input. We used the species tree obtained with OrthoFinder, which in turn required additional conversion with the software Tree from the python-based package ete3 (Huerta-Cepas et al. 2016). For each orthogroup identified in OrthoFinder we tested five different models, reflecting the evolution of the gene family. The null model (Global rate model) assumed uniform rate of gene gain/loss for all branches in the provided species tree. Alternative models were specified as following:

1) To detect gene families changes specific to *V. cardui*, we specified distinct branch rate in *V. cardui* with all the other branches evolving with single background rate

2) In the second model the *V. cardui* and *D. plexippus* branches shared the common rate of change, which allowed us to find gene family expansions common for these migratory butterflies, but absent in the other taxa.

3) The third branch specific rate included the *Vanessa* genus branch (both V. cardui and V. tameamea) with one rate compared to the background rate.

Each model was run twice and the replicate with highest likelihood was used for model comparison.

The null model and each alternative model were compared using Aikaike's Information Criterion (AIC) calculated as 2K-2logL where K is the number of parameters and logL is the logarithm of the likelihood of the model. The orthogroups where the alternative models in BadiRate inferred gene gains or losses >0 and had lower AIC was used for further analyses: functional assignment enrichment and spatial distribution along the genome. The program was partly run with a modified version of the R-package BadiRateR (https://palfalvi.github.io/badirater/articles/badirater.html) and custom scripts.

We investigated the location of genes belonging to orthogroups identified in BadiRate and visualized their distribution using custom bash and python scripts (available on GitHub). We used information from the annotation gff file and names of the genes from BadiRate

Gene ontology enrichment

We tested for enrichment of functional categories in the gene sets of interest with the Bioconductor package topGO version 2.44.0 (Alexa and Rahnenfuhrer 2021) in R version 4.1.0 (R Core Team, 2013) using a custom database 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 and the default algorithm ("weight01") accounting for the hierarchical structure of the GO-terms (Alexa et al. 2006). This means that the resulting tests are not independent and correcting for multiple testing might not be motivated, however to keep a conservative approach we adjusted the p-values with Benjamini-Hochberg's method of multiple test correction (p.adjust(x, method = "fdr")).

Comparative analysis of genes associated with migration

We investigated the presence or absence of previously described genes associated to migration in the V. cardui geneset. This was performed by doing a nucleotide BLAST search of the V. cardui geneset to a list of genes associated to biological functions of interest for migration in the Monarch butterfly (Danaus plexippus). The gene sets were accessed 210611 from MonarchBase http://monarch.umassmed.edu/group.html together with a reference fasta file Dp_geneset_OGS2_pep.fasta.gz from http://monarch.umassmed.edu/resource.html. The nucleotide sequences for the migratory gene set were extracted from reference with a custom script resulting in 588 Monarch genes of interest to use as BLAST database (ref BLAST).

Patterns of recombination

Recombination rate analysis

The linkage maps were cleaned and rearranged so that any markers mapping to different chromosomes were removed and the maps arranged in ascending map position with custom R script and finally markers deviating from ascending order were removed with R-script from (https://github.com/tsackton/linked-selection) (Sackton et al. 2014). Recombination rate was estimated with the R-package MareyMap for each marker using linear regression in 2 Mb sliding windows containing more than 2 markers (Rezvoy et al. 2007). A custom R-script transformed the recombination rate per marker to recombination rate in 2 Mb windows across the genome.

Correlations between genomic features

Correlation test between different genomic features was performed with cor.test in R using Spearman's rank correlation, after testing for normality with Shapiro-Wilk normality test (R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/). We used lm in base R to explore the relationships in a linear model and lmer in the R-package lme4 (ref) for a mixed linear model after scaling and centering of the explanatory variables. To further analyse the association between gene gains and other genomic features the gene gains was classified into two categories, with or without gained genes, for the window sizes 100kb and 2 Mb. The difference in window mean was tested with non-parametric Wilcoxon test implemented in the package rstatix (Alboukadel Kassambara (2021). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.0. https://CRAN.R-project.org/package=rstatix). All the tests were performed in two versions, one including all windows along the genome, the other excluding the W-chromosome. We used the R-package ggplot2 for visualisation (Wickham 2009).

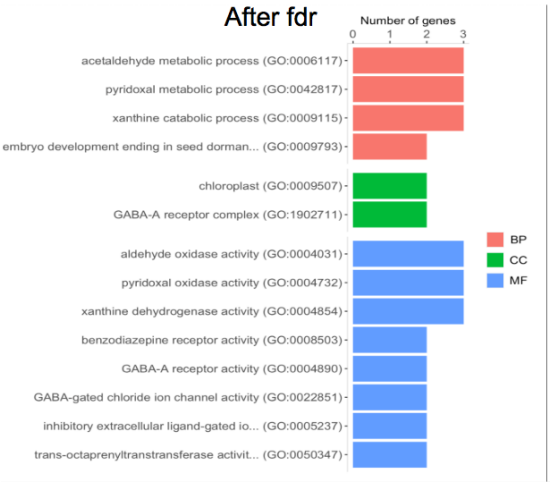
Window-based analysis

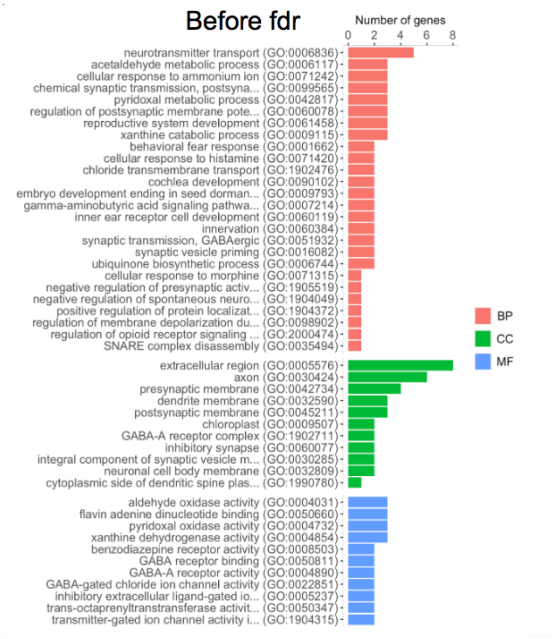
Supplementary

1. Marey and rec rate per chromosome
2. Table rec rates per chrom
3. Table with general stats for gene family analysis

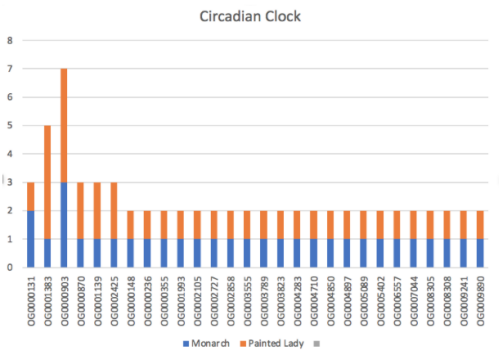
4. Monarch model

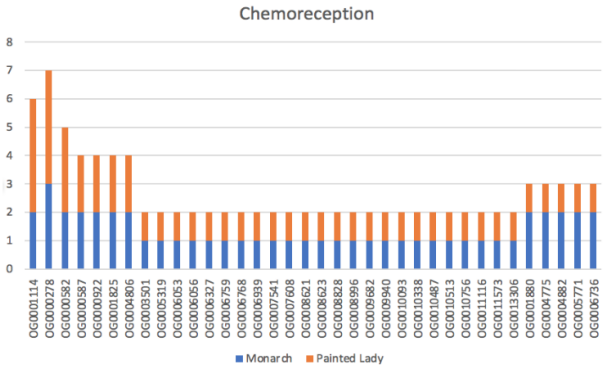
Number of orthogroups: 39
Nr of genes: 196
Enriched GO-terms:

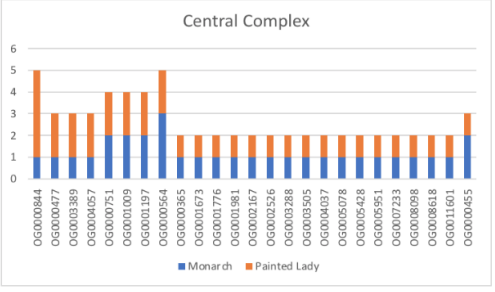


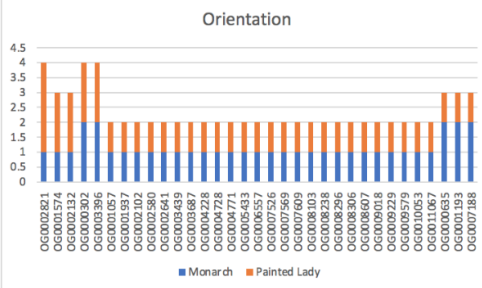


5. Monarch candidate genes









	Number of elements	Length occupied	Percentage of sequence
SINEs:	115094	21196552 bp	4.92%
LINEs:	36820	9869881	2.29%
LTR elements:	11547	5335487	1.24%
DNA elements:	52850	7889282 bp	1.83%
Unclassified:	292186	37717554	8.76%
Total interspersed repeats		82008756 bp	19.05%
Small RNA	66510	13989434	3.25%
Satellites	132	9117	
Simple repeats	144667	6408573 bp	1.49%
Low complexity	23725	1124596 bp	0.26%

References Vancouver reference style

J.Huerta-Cepas, D.Szklarczyk, D.Heller, A.Hernández-Plaza, S.K.Forslund, H.Cook, D.R.Mende, I.Letunic, T.Rattei, L.J.Jensen, C. von Mering, P.Bork ork
eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses.

(Genome research style)

Alexa A, Rahnenfuhrer J. 2021. *topGO: Enrichment Analysis for Gene Ontology*. Bioconductor version: Release (3.13) https://bioconductor.org/packages/topGO/ (Accessed August 24, 2021).

Alexa A, Rahnenführer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**: 1600–1607.

Aljanabi S. 1997. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Res* **25**: 4692–4693.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**: 238.

Espeland M, Breinholt J, Willmott KR, Warren AD, Vila R, Toussaint EFA, Maunsell SC, Aduse-Poku K, Talavera G, Eastwood R, et al. 2018. A Comprehensive and Dated Phylogenomic Analysis of Butterflies. *Curr Biol* **28**: 770-778.e5.

Herrera S, Reyes-Herrera PH, Shank TM. 2015. Predicting RAD-seq Marker Numbers across the Eukaryotic Tree of Life. *Genome Biol Evol* **7**: 3207–3225.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* **33**: 1635–1638.

Librado P, Vieira FG, Rozas J. 2012. BadiRate: estimating family turnover rates by likelihood-based methods. *Bioinformatics* **28**: 279–281.

Rezvoy C, Charif D, Gueguen L, Marais GAB. 2007. MareyMap: an R-based tool with graphical interface for estimating recombination rates. *Bioinformatics* **23**: 2188–2189.

Sackton TB, Corbett-Detig RB, Nagaraju J, Vaishna L, Arunkumar KP, Hartl DL. 2014. Positive Selection Drives Faster-Z Evolution In Silkmoths: Faster-Z Evolution In Silkmoths. *Evolution* n/a-n/a.

Wickham H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York https://www.springer.com/gp/book/9780387981413 (Accessed August 24, 2021).