# Table of contents

[**Table of contents**](#_ofw2l3ckhr08) **1**

[**Abstract**](#_nmdcflie2eds) **2**

[Keywords](#_vxdo9ay9yaes) 2

[**Background**](#_t70fsdekxwv) **2**

[**Results**](#_xsii4sebg2jz) **4**

[Genome statistics](#_ofo1e6nb1m70) 4

[Synteny](#_fkkjaaa7p6mw) 5

[Gene family evolution](#_3tx5somlohvf) 6

[Location of expanded families along the genome.](#_z8cygwitt9sv) 8

[Identification of candidate genes of interest in the monarch.](#_prdzj11gqp25) 8

[Patterns of recombination](#_rxx8yikkz2ni) 8

[Chromosome level](#_5pf4z29qmvb2) 8

[Regional variation](#_u283l5vgps6q) 9

[Spatial correlation between gene duplications and genomic features](#_xthqq3a4i4s2) 10

[**Discussion**](#_dxjj682ttjk6) **14**

[Linkage map improves analysis of genome evolution](#_uguqnifj1lbs) 14

[Adaptation to migration](#_r32wm5nvn894) 15

[Evolution of high adaptability](#_1muzci9d7mdd) 16

[Role of recombination in genome evolution](#_3u31swi6s7c6) 16

[Recombination landscape in V.cardui](#_dy6s2vu6i57z) 17

[Gene duplication and genomic architecture](#_rt4tvnot6jwj) 18

[Evolution of sex-determination system](#_3uxqz9oa4ys5) 19

[Conclusions](#_5a6s5no4aq0t) 19

[**Methods**](#_oulz40hkhizt) **20**

[Linkage map](#_tqqlyb3uarto) 20

[Sampling and DNA-extraction](#_madoc62fk0cz) 20

[Data processing](#_hd01vnih3p2z) 20

[Construction of linkage map](#_fjb81j1dzw8f) 21

[Genome annotation and whole genome statistics](#_rx4i90fcpfs1) 22

[Genome assembly statistics](#_p2jp79gdehlt) 22

[Gene and repeat annotation](#_64gfqf47jf7l) 22

[Gene family evolution](#_zg1ny7az0y4a) 23

[Gene ontology enrichment](#_goin7r8ptchk) 24

[Comparative analysis of genes associated with migration](#_6e7ajyrzqyun) 24

[Patterns of recombination](#_ympt1vbdqrtk) 25

[Recombination rate analysis](#_bodlqwh9vmeg) 25

[Window-based analysis](#_5bpkj1du205z) 25

[Correlations between genomic features](#_qwhk1gaxsfxk) 25

[**Supplementary**](#_7wnt5zpge4yz) **28**

[**References ​​**](#_zchoynj625et) **30**

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

## Genome statistics

We verified the existing assembly from the Darwin Tree of Life project with a pedigree based linkage map. The total map length is 1375 cM, and the maps were constructed with 1696 markers with average marker density of 3.95 markers/Mb. Genome occupies total length of 424Mb and consists of 30 autosomes and the sex chromosomes Z and W (summarized in Table 1.). We didn’t observe significant differences in basic statistics between linkage map corrected assembly and reference DToL assembly. Large N50 and high BUSCO scores allow us to suggest that chromosomes are nearly complete. GC content equals 33.41%

Total length of sequences marked as repeats slightly exceeded 100Mbp (23.38%): LINEs and SINE are among the most represented repeat classes. We identified 14,957 protein-coding genes, from which 12,098 had functional annotation.

|  |  |  |
| --- | --- | --- |
| **Genome statistics** | |  |
|  | Total sequence length | 424,577,752 |
|  | Number of chromosomes | 32 |
|  | Scaffold N50 | 14,615,999 |
|  | GC content | 33.41% |
| **Repeat statistics** | |  |
|  | SINEs | 4.96 % |
|  | LINEs | 2.27 % |
|  | LTR elements | 1.21 % |
|  | DNA elements | 1.81 % |
|  | Simple repeats | 1.46 % |
|  | Total repeats | 23.40% |
| **Genome annotation** | |  |
|  | BUSCO genes (arthropoda) | 97% |
|  | High quality gene models | 14,957 |
|  | Genes with functional annotation | 12,098 |

Table 1. Genome statistics of *Vanessa cardui*

## 

## Synteny

We assessed collinearity of the *V.cardui* genome using annotated genes and synteny alignment on two levels of genetic divergence: we first compared the *V.cardui* genome to the genome of the Silk moth *(Bombyx mori)* and on a closer evolutionary scale we performed a comparison with another species from the 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 in both comparisons. Notable exception from this pattern is the W-chromosome, which is not assembled 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 several 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)](https://www.zotero.org/google-docs/?nEl1CT). OrthoFinder inferred 86 duplications on the common *Vanessa* genus branch and 1125 duplications on the tip branch to *V. cardui* compared to 6218 in *V. tameamea* (Figure A).

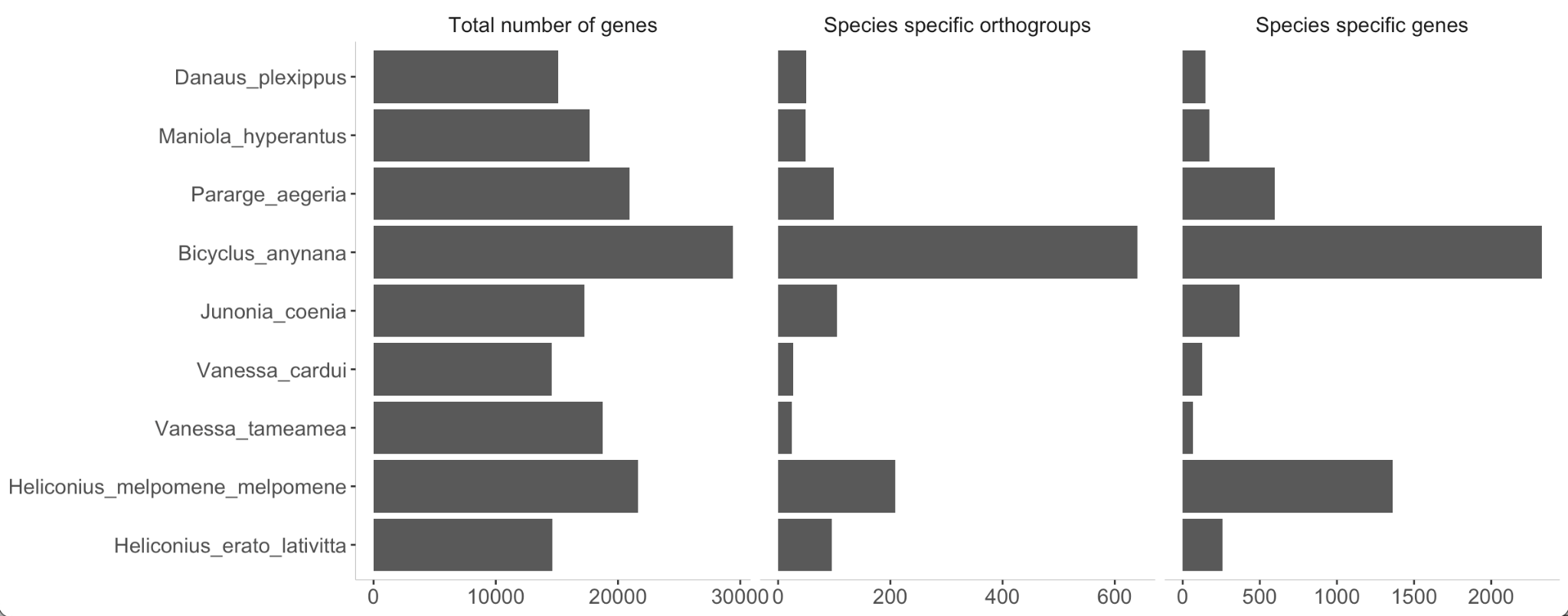
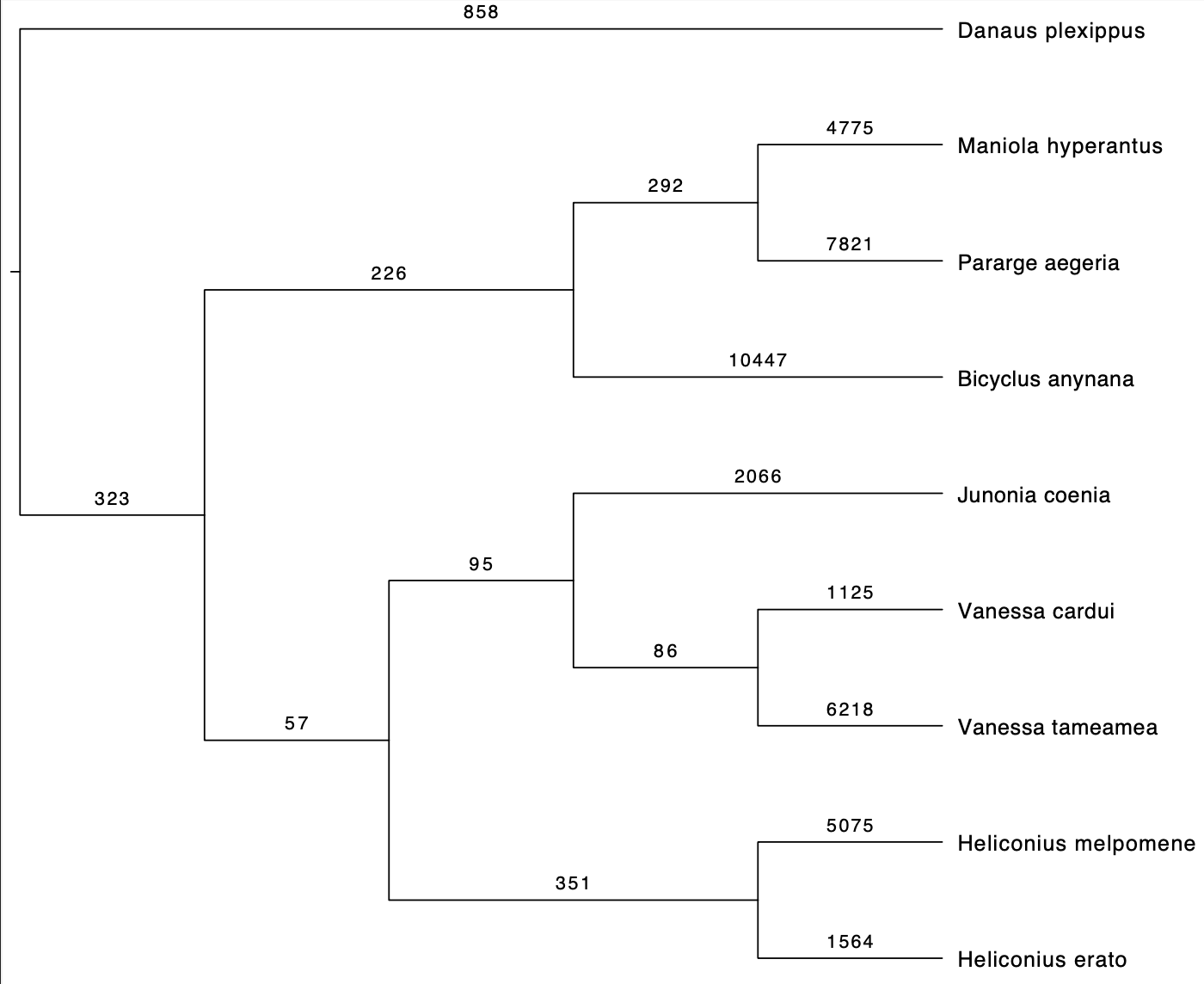


Figure A. Gene family evolution on Nymphalid species tree inferred by OrthoFinder. The number on the branches shows duplications. The barplots show the total number of genes included, species-specific orthogroups and number of genes in species-specific orthogroups.

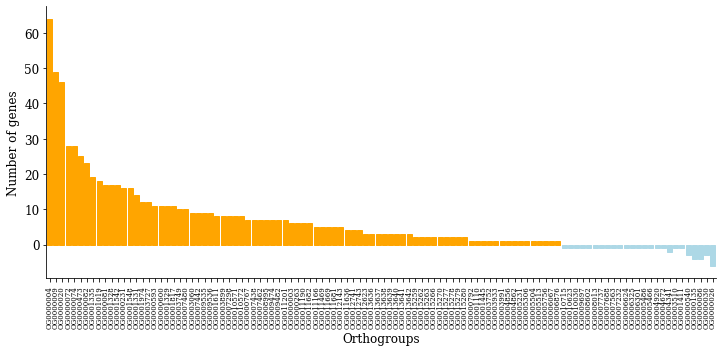
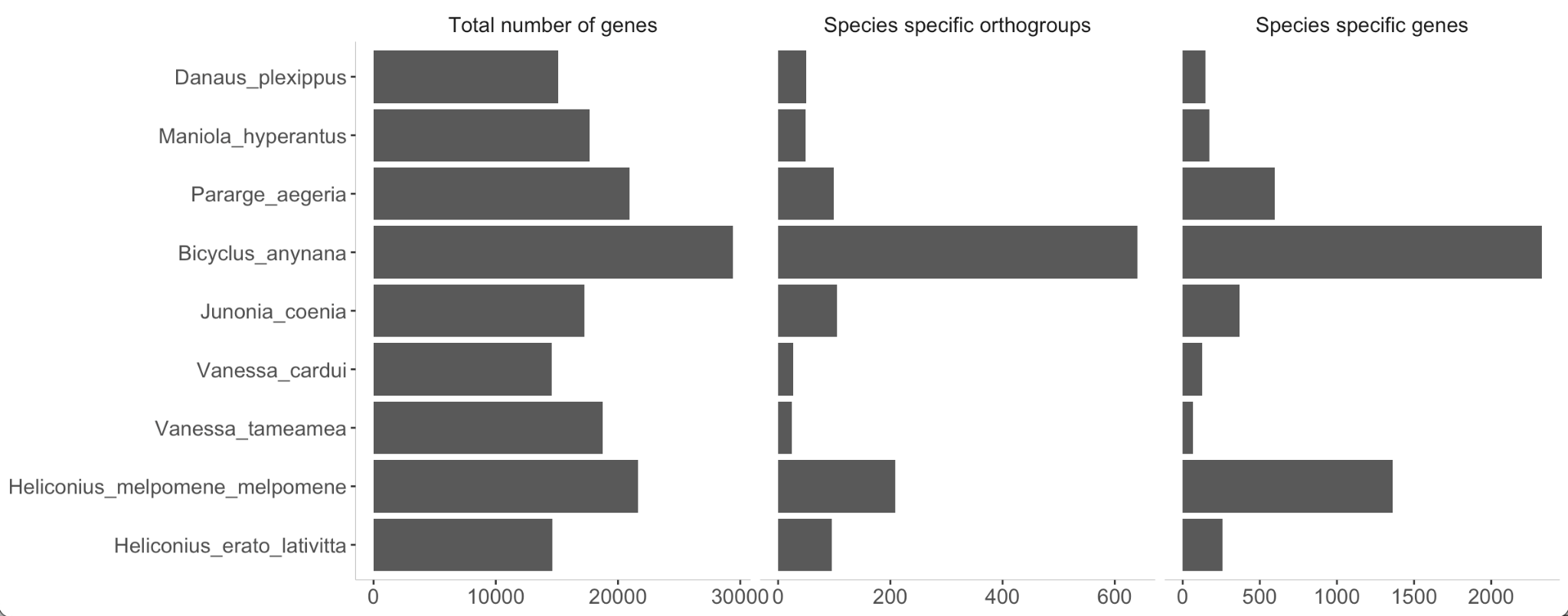
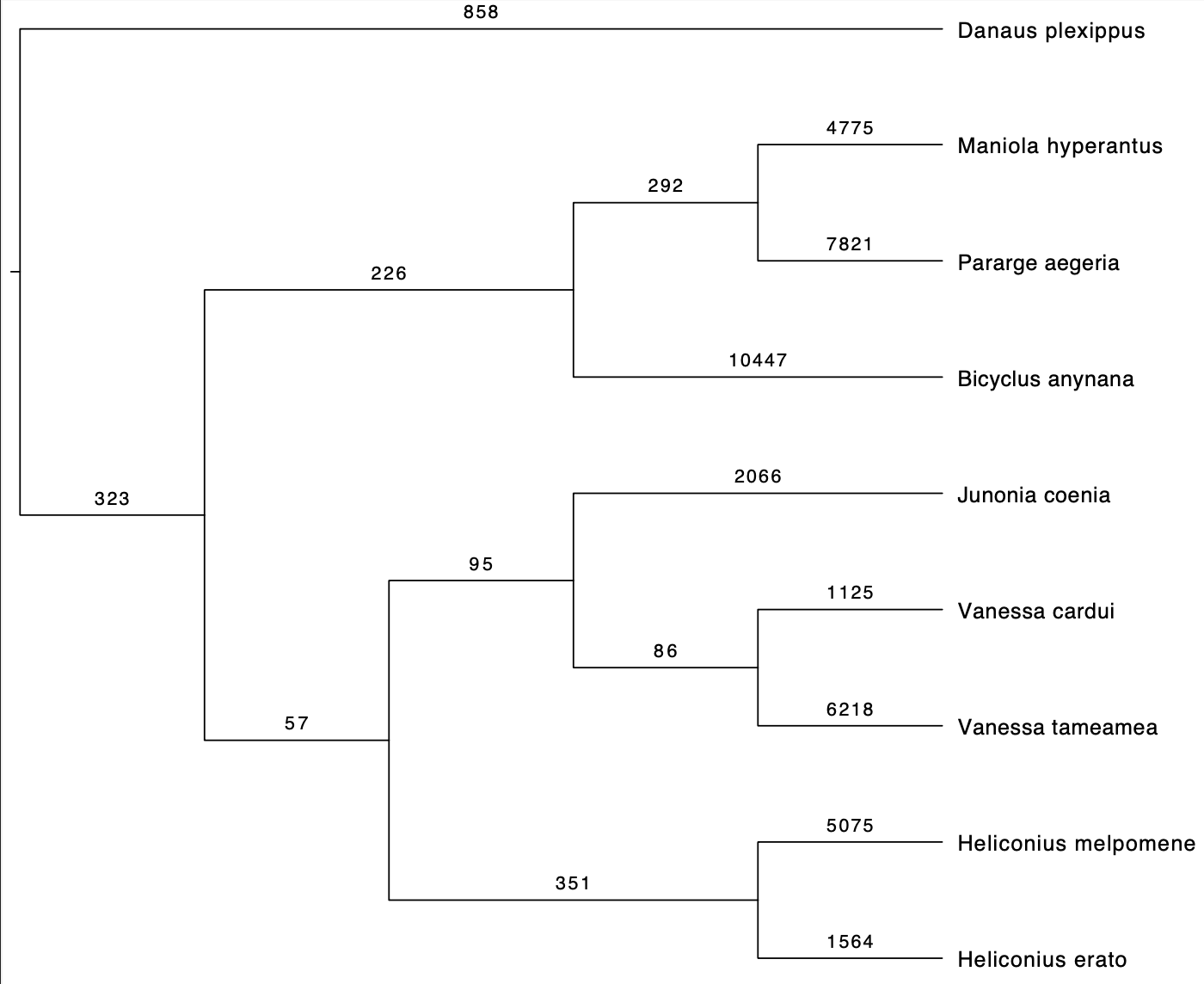


Figure A. (Option 2)

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).

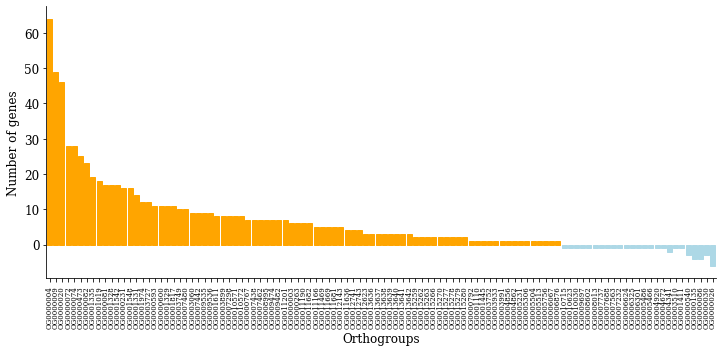
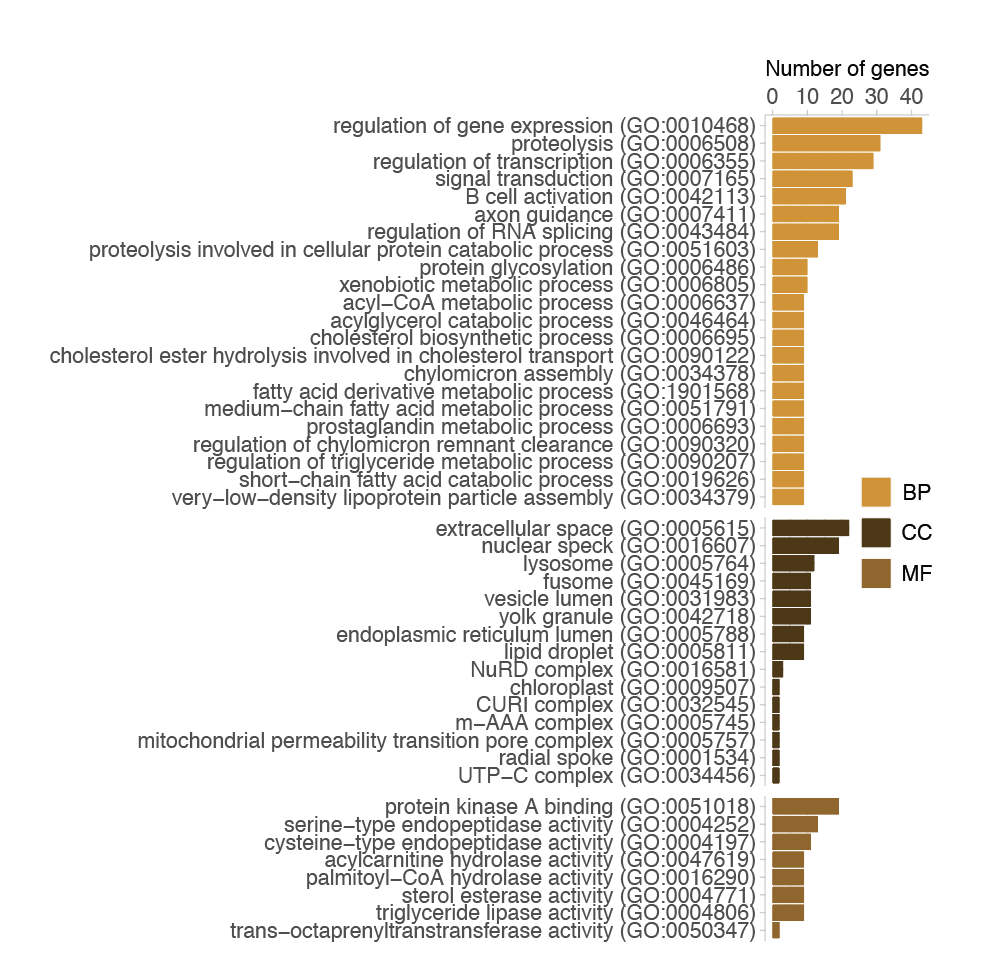
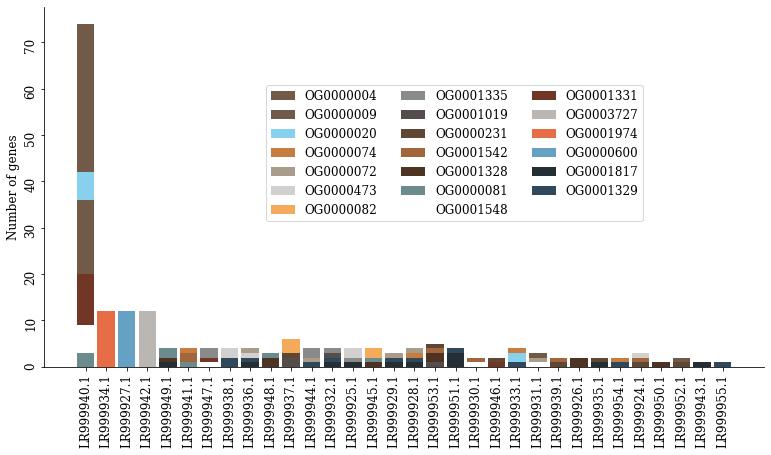
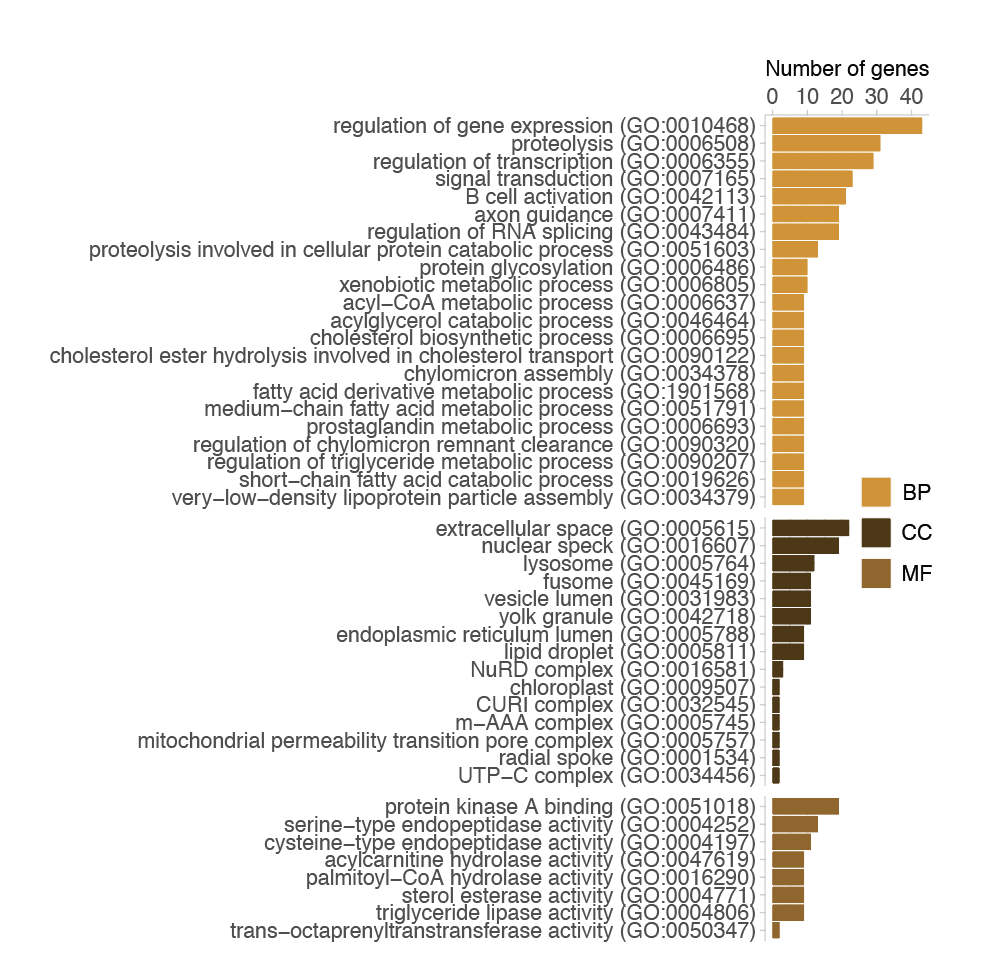


Figure B. A) Enriched gene ontology (GO) terms associated to the genes gained in the V. cardui branch, with p-value < 0.05 after FDR-correction. The bars show the number of genes associated to each GO-term. The different GO-categories are biological processes (BP), cellular compartment (CC) and molecular function (MF). B) Orthogroups that experienced gene gain or loss according to 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.

Figure B. (Option 2)

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.

Among 588 genes involved in migration in Monarch butterfly (*Danaus plexippus*), only 259 had orthologs in *Vanessa cardui* genome. Common genes involved in orientation would in 33 orthogroups, which included 38 genes in Monarch and 39 Painted Lady. Among circadian clock associated genes (28 common orthogroups) number of *V.cardui* genes was slightly reduced (31, compared to 37 in Monarch). For other groups of candidate genes, number of genes was nearly identical for both species (central complex: 60 orthogroups / 62 genes in *D.plexippus* / 70 genes in *V.cardui*, chemoreception: 36/49/48, sensory input: 86/104/108, RNA interference: 18/22/22). We didn’t find any expansions or contractions of genes families from the orthogroups above in *Vanessa*-specific branch model.

#### Location of expanded gene families along the genome

Chromosome level assembly enables us to look at spatial distribution of gene duplications along the *Vanessa cardui* genome. On the chromosome level we see an uneven pattern of accumulation of genes (species-specific gene gains considered, distribution of the most abundant orthogroups visualized on Figure B C)). While on average chromosome harbours 23.7 genes, this number varies dramatically from 6 to 175 (std +- 29.5). W-chromosome is a clear outlier with the maximum number of gained genes: 175 genes from orthogroups. It is followed by three more chromosomes (3, 10, 18) containing 47, 48, 44 genes (each belonging to a unique orthogroup). The rest of the chromosomes have a number of gained genes close to average value.

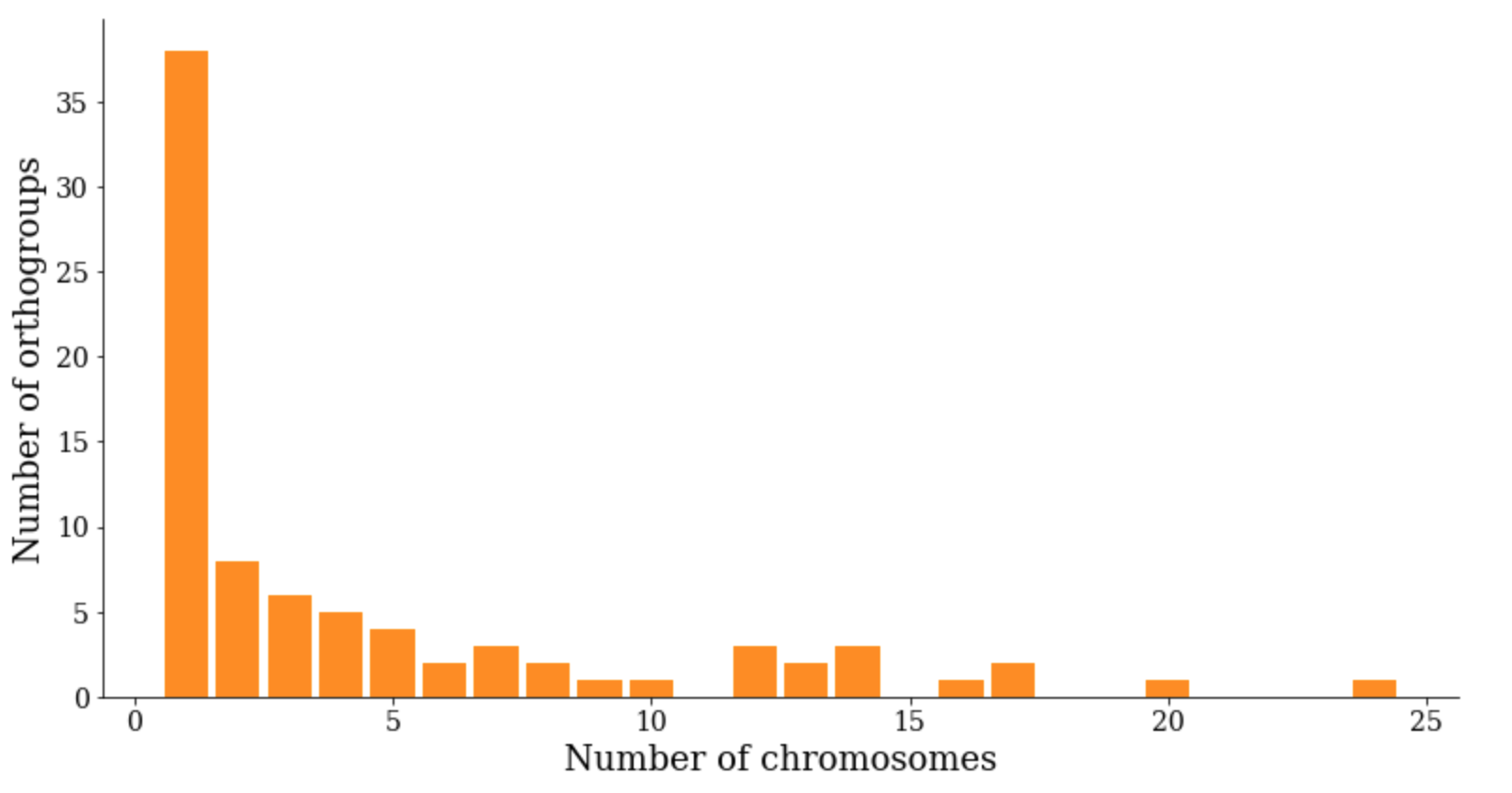


Figure. Number of orthogroup occurrences per chromosome count

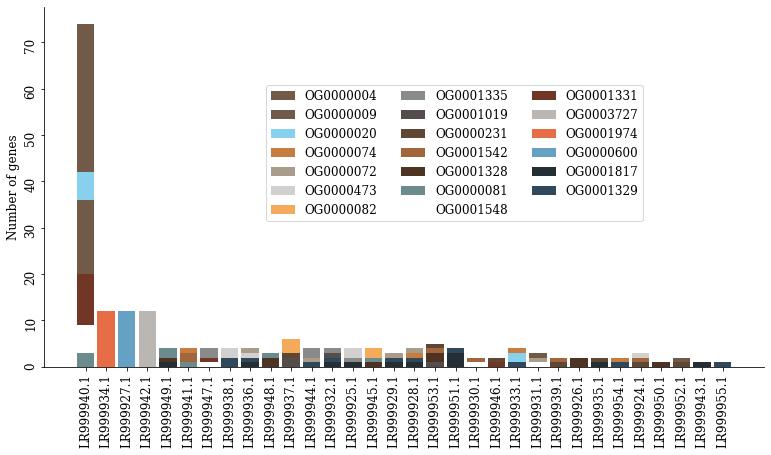
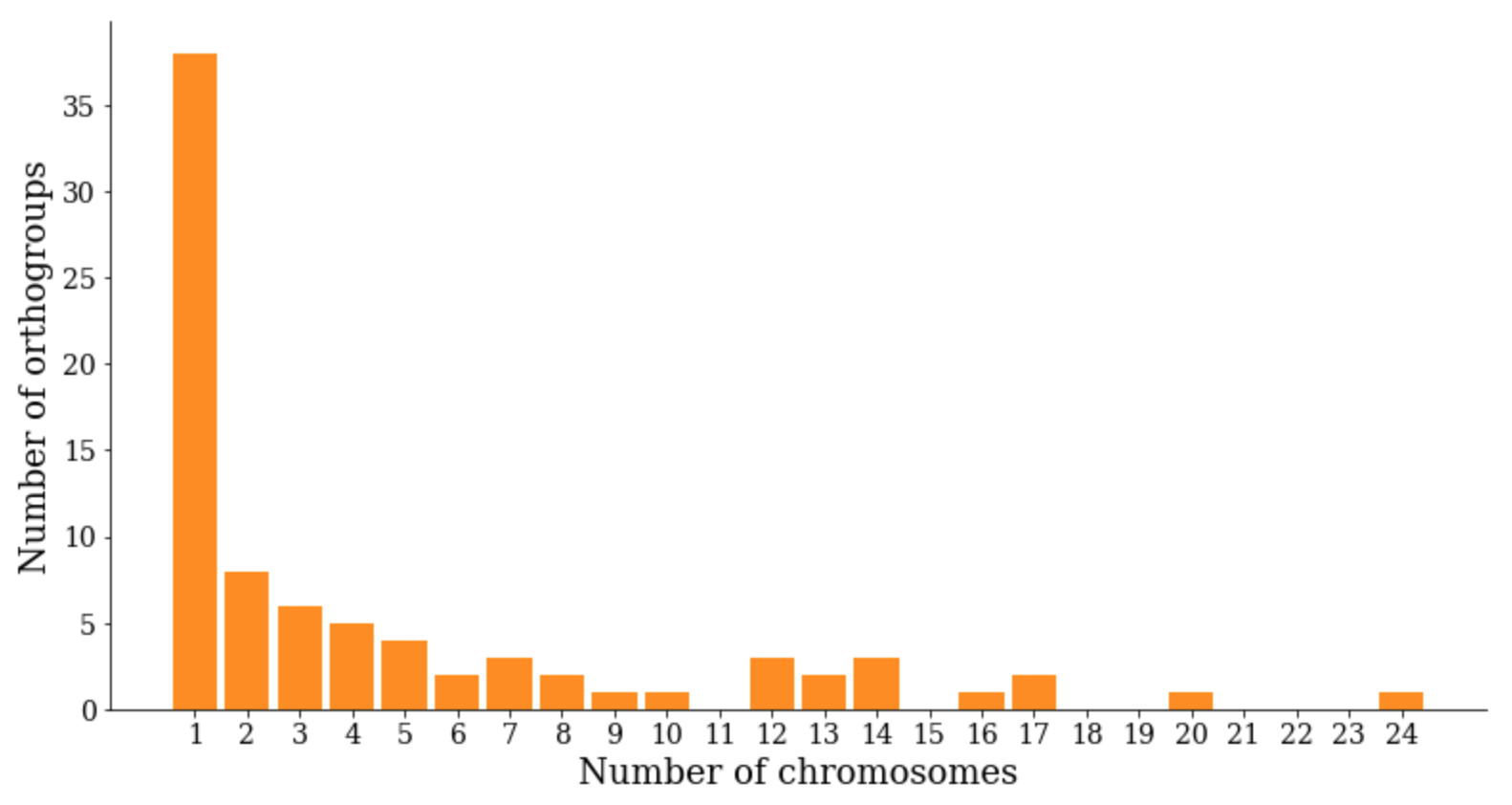


Figure B B (Option 2). Number of orthogroup occurrences per chromosome count

Spatial distribution orthogroups follow two distinct scenarios: a) all the gene duplications happen located in the same chromosome and form clusters, b) genes belonging to the extended orthogroup distributed between several chromosomes (how many?). We further performed fine-scale visualization of gene gains distributions, which we discuss below in connection to distribution of other genomic features (see chapter window-based analysis, Suppl. Figure (with all chromosomes)).

### 

## Patterns of recombination

Construction of the linkage map allowed us to obtain the estimate of the recombination rate in *Vanessa cardui*. The average genome wide recombination rate (calculated as the map length divided by genome size) is 3.19 cM/Mb with variation between chromosomes 1.75-5.34 cM/Mb. W chromosome slightly lowers global estimate, while excluding the W chromosome the global recombination rate is estimated as 3.42(+/-0.72)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.

### Chromosome level

We investigated factors which may influence variation of recombination rate between chromosomes (Figure C) and regionally on the level of the same chromosome (Figure D). For this we considered length of the chromosomes, repeat and gene content.

On the chromosome level…. 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. 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 ).Gene number is positively correlated to chromosome length.

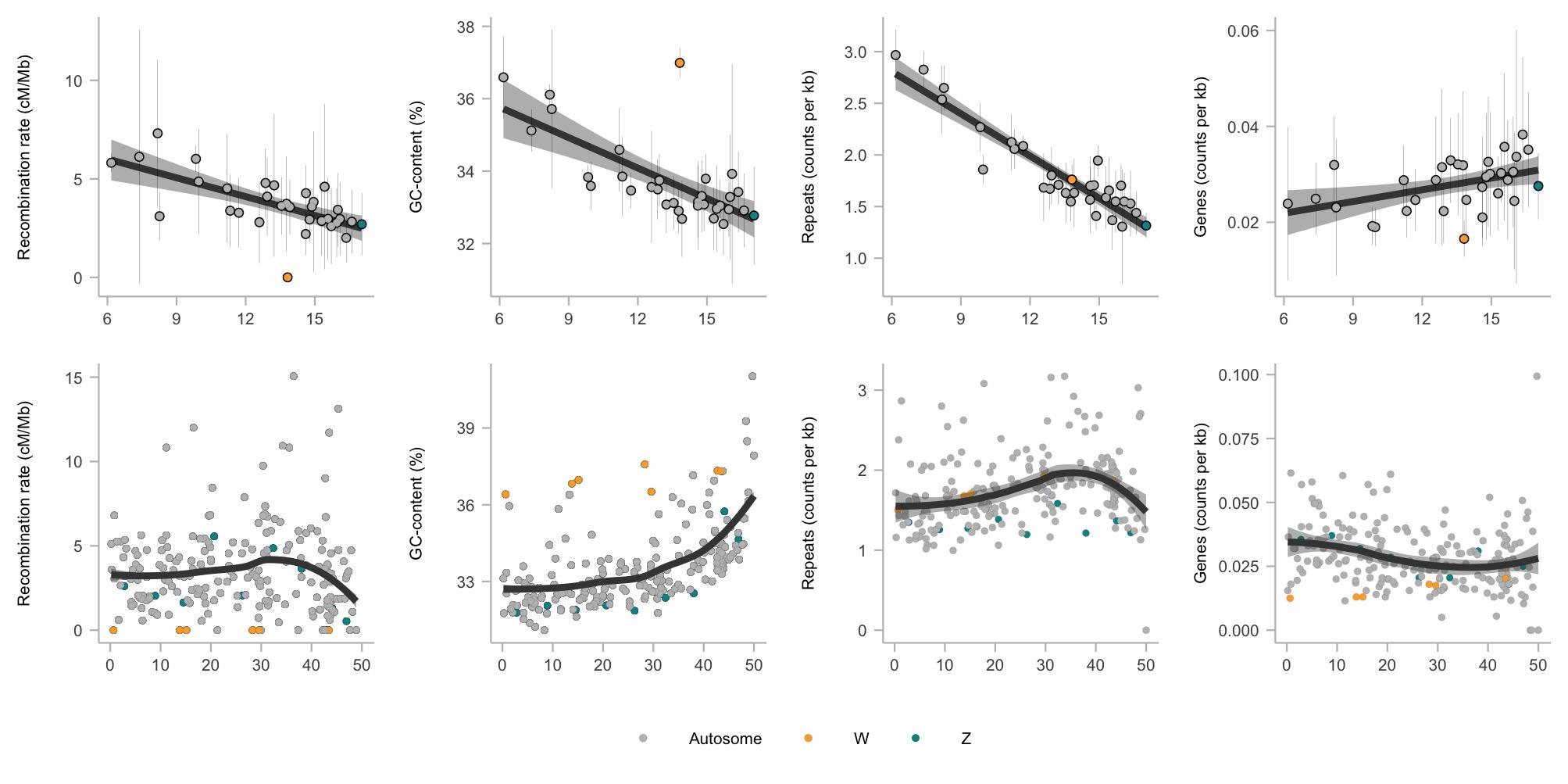


Figure C and D. Correlation between chromosome length and genomic features. Regional distribution

At the second step we considered regional variation in recombination rate within chromosomes, by applying a window-based approach. Taking genomic windows into consideration (non-overlapping window-based linear regression, window size 2 Mb) lead to slight change in recombination rate estimates. The global recombination rate average across all windows in the genome was estimated at a slightly higher level of 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.

### Regional variation

Window-based estimates of the recombination rate made possible to evaluate regional variation in recombination rate and analysed chromosomes jointly. Despite expectation for holocentric chromosomes, the recombination rate is not uniformly distributed along the chromosome: visual inspection (Figure D a) shows somewhat lower recombination rate in the centre of the chromosome and it’s significant reduction in the terminal regions. Statistical evaluation was performed to test this observation. We binned the markers in five distance intervals from the centre of the chromosome (combining information from all chromosomes). Reduction of recombination rate in flanking regions (the last bin) was statistically significant compared to the centre and terminal regions (stats). Decrease of the recombination rate in the centre of the chromosome wasn’t statistically significant (stats).

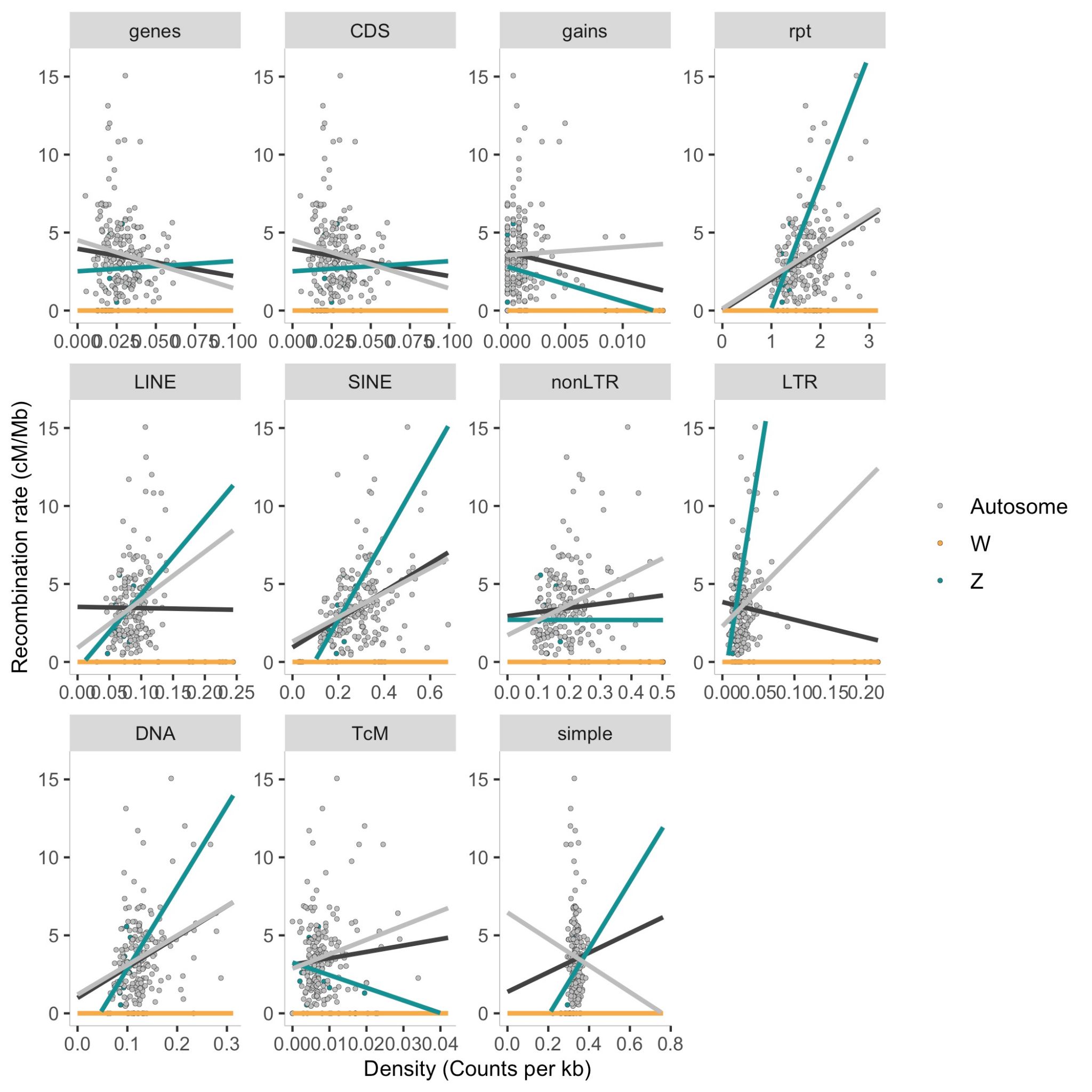
We investigated factors influencing variation of recombination rate along the chromosomes. First, we tested if some the genomic features of consideration demonstrate similar patterns of distribution, secondly we obtained a linear model to test correlation of these patterns to recombination rate. Analogous to the whole genome analysis we tested the influence of gene and repeat content. We observed that the regional distribution of total repeat content follows the same pattern as the recombination rate (Figure D c). Looking at each of the repeat classes separately, we mostly observe the same pattern, the only exception of LTR (long terminal repeat) and the Tc1/mariner (superfamily of interspersed repeats DNA (Class II) transposons). 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).

### Spatial correlation between gene duplications and genomic features

We investigated factors which may have influenced the rates of gene duplication in Vanessa cardui: The relationship between gene gains and genomic features was explored on in 2Mb window size, including the recombination 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 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). The recombination rate does not appear to have 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.

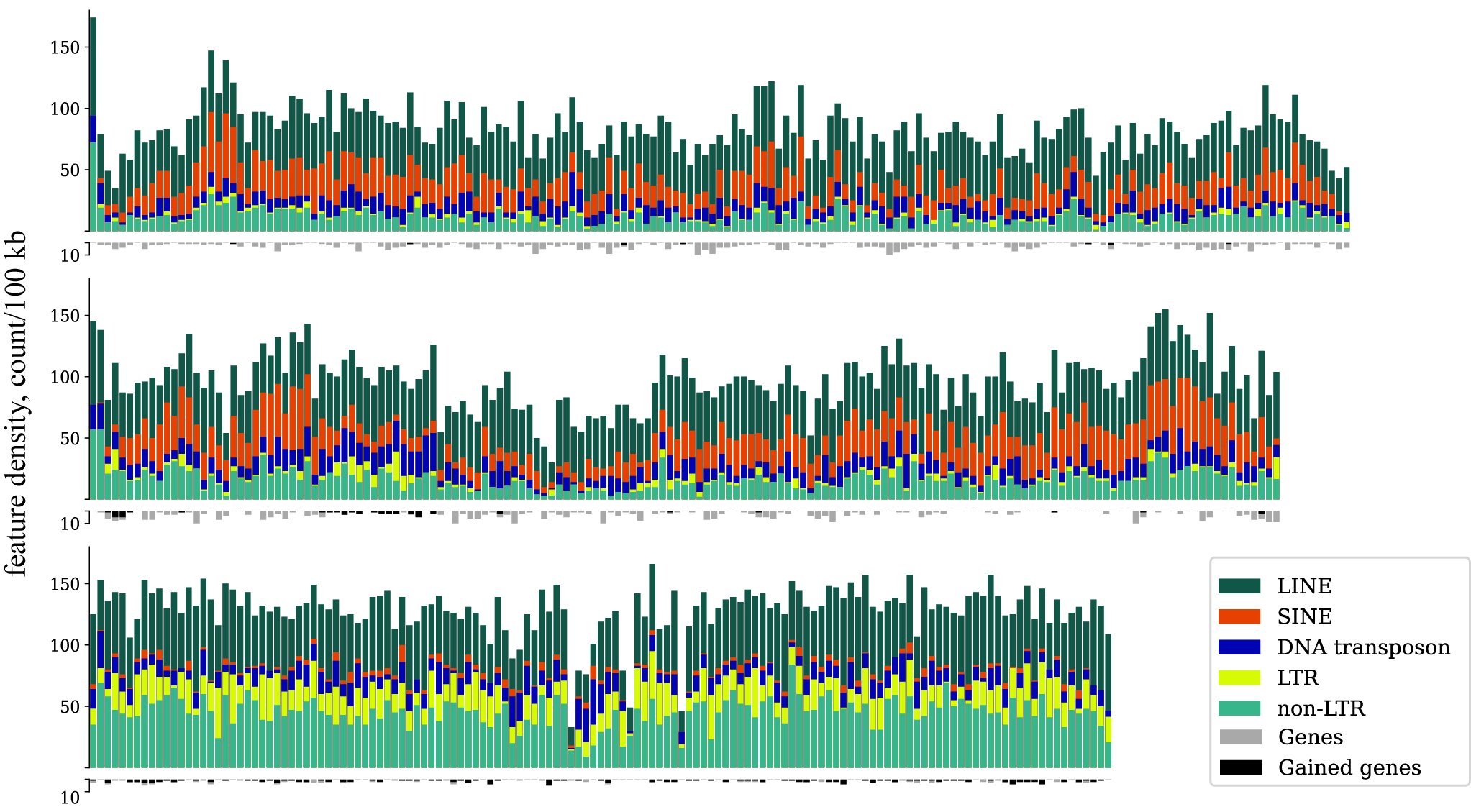
We then compared the features on a detailed level in 100 kb windows within a linear model, and now there is a small effect of gene density on the number of genes gained on the autosomes. Other significant effects on the number of gained genes on the autosomes are a negative effect of SINEs and a positive 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*



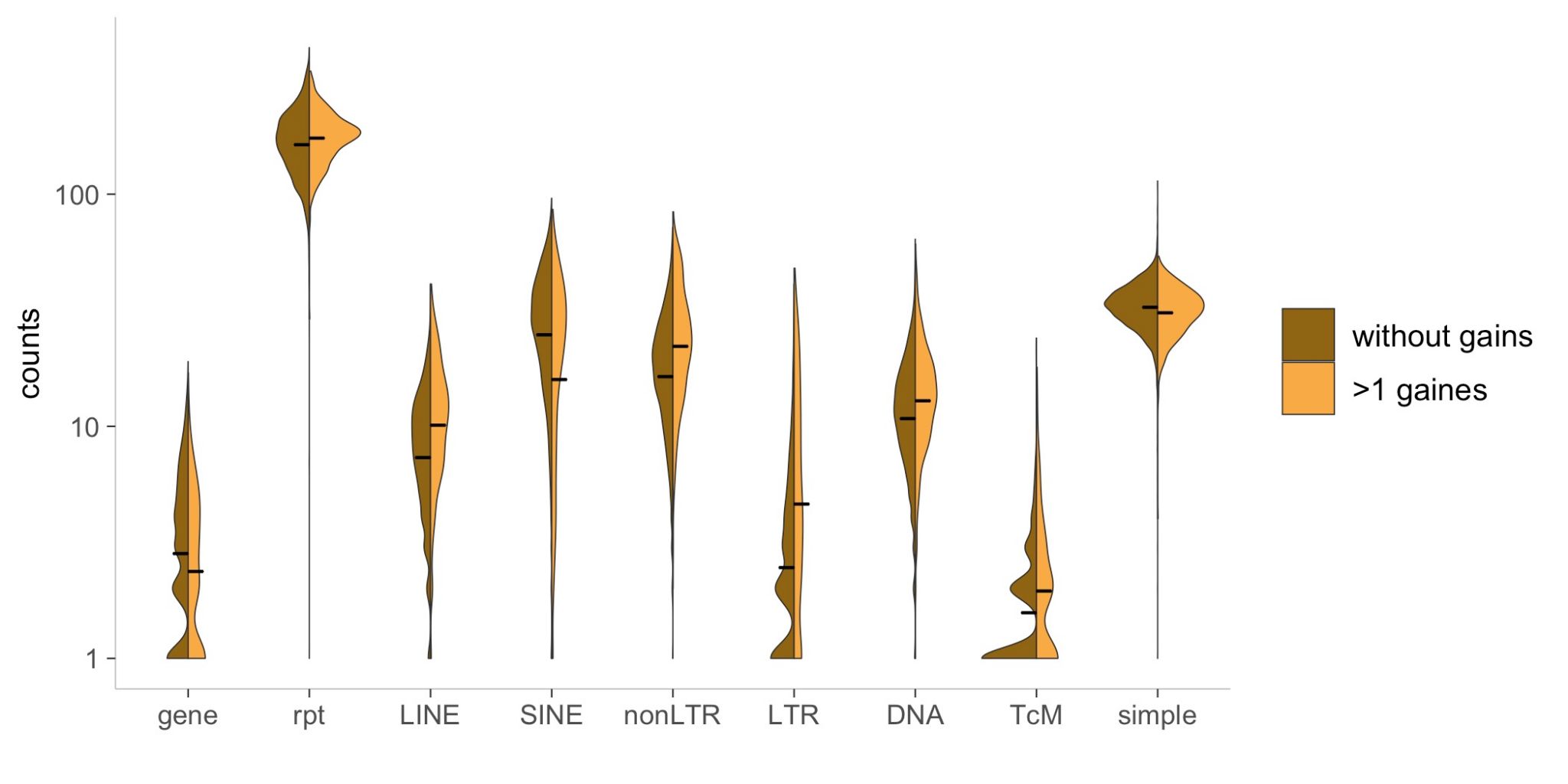
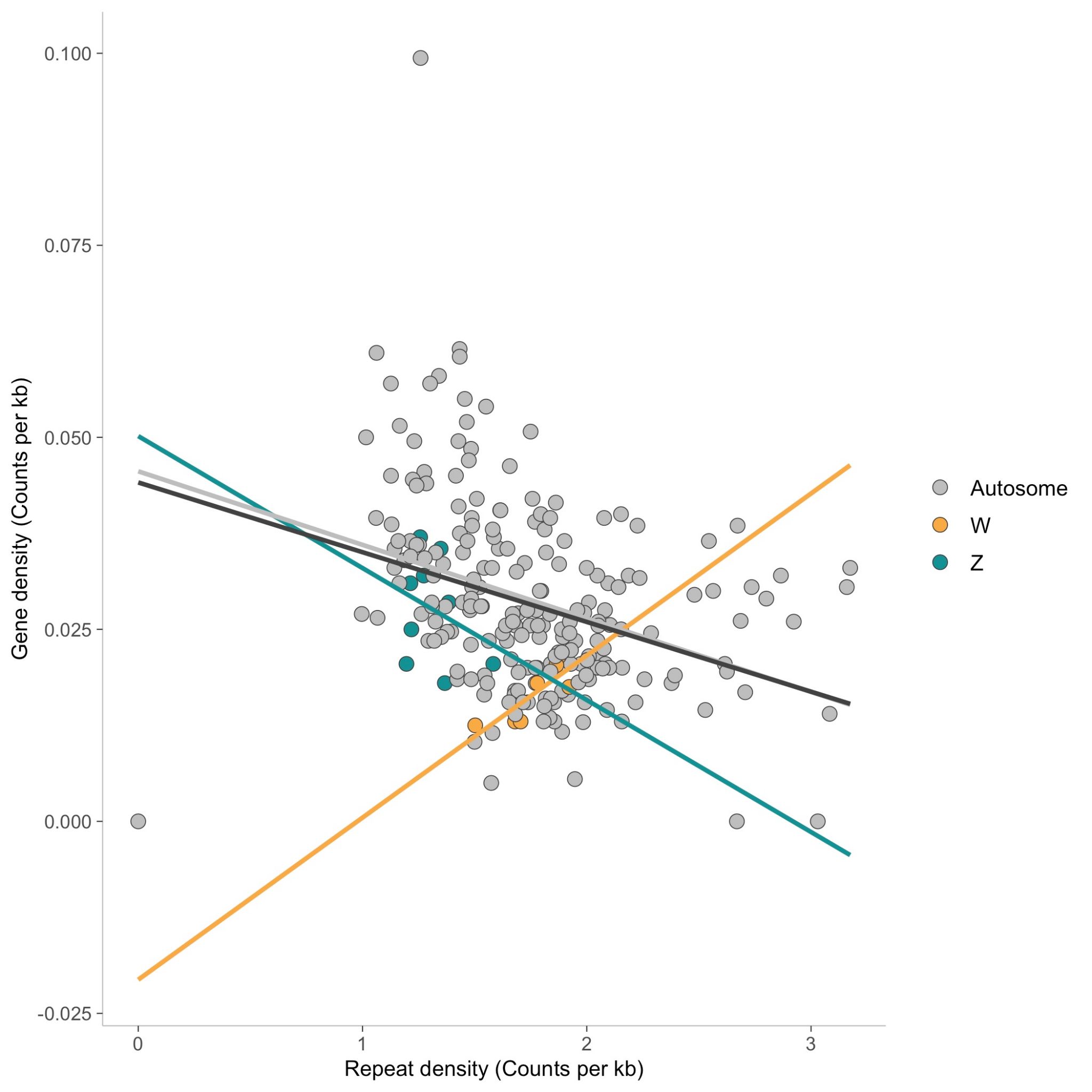


Figure F (preliminary). Window-based analysis: distribution of repeat classes, genes and genes belonging to expanded gene families along the chromosomes. Figure G. Correlation between genes and repeats. Figure H. Violin plot of gene gains vs no gene gains in different genomic features.

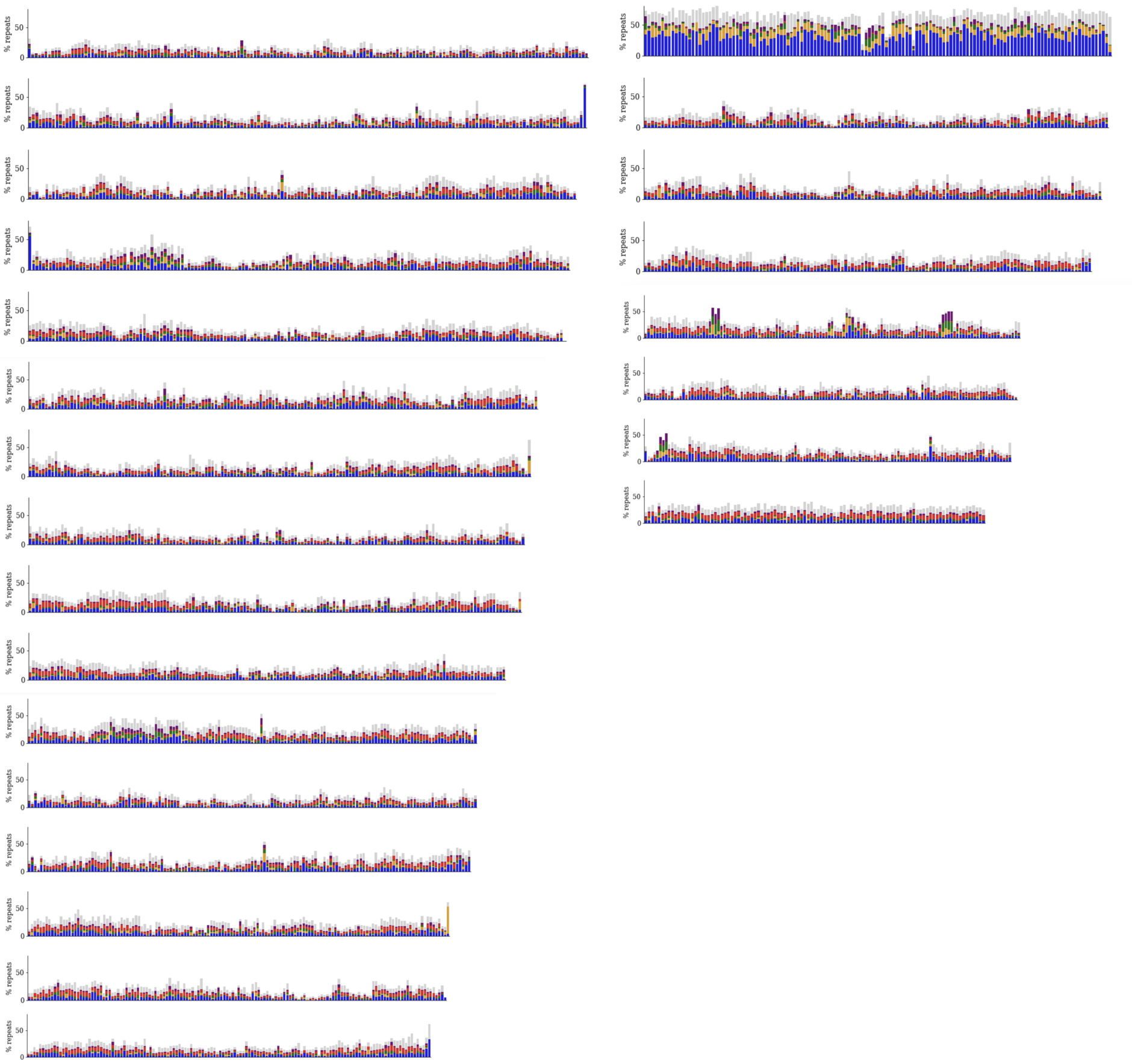


Figure F (version 2).

# Discussion

In this study we improve genome assembly for further analysis, analyse interplay of selection, recombination and selfish genetic elements shaping the genome of *Vanessa cardui* and provide first insights on adaptation to migratory lifestyle. We are able to make broad conclusions applicable to the entire Lepidoptera.

## Linkage map improves analysis of genome evolution

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: using long-read sequencing, scaffolding with Hi-C and additional verification using linkage map. Construction of linkage map is of particular importance for this paper, since it gives access to: a) verifies contiguity of chromosome level assembly, b) gives new insights to pattern of recombination and opens up possibility for all consecutive analysis of interaction between different evolutionary forces, d) allows reliable analysis of spatial distribution of genes and repeats (TE).

High quality genome assembly allows the first comparative look at the Vanessa cardui genome within Lepidoptera in general and Nymphalid clade in particular. Vanessa cardui is extraordinary in its natural history traits as it: performs one of the longest migrations known to insect species, cosmopolitan and has high adaptability to a variety of environmental conditions and host plants.THis combination of trait also results in high levels of heterozygosity and effective population size. In some cases phenotypic traits may manifest themselves already on the level of the basic genomic features. Vanessa has average genome size (Table A) and presumably ancestral karyotype and high synteny (Figure H). Vanessa genome is characterized by a relatively small number of genes, repeats and moderate recombination rate. Analysis of species-specific gene families shows the values on the lower end of the spectrum, while overall number of duplications on Nymphalid tree is very variable (Figure A).

Another speciality of the presented genome assembly is presence of the W chromosome. This chromosome presents a specific challenge for assembly projects that is not present in the majority of available Lepidoptera genomes. We observe very little connection with other known chromosomes from Bombyx and Heliconius, which confirms identification of the W chromosome. We discuss specific traits of w chromosome and Vanessa sex deteremination system below.

## Adaptation to migration

Vanessa cardui is a striking model organism for studying butterfly migration, but as of now genomic features underlying this behaviour were only characterized for a handful of species and nothing is known about Vanessa in particular. Genomic analysis of such complex phenotypes as a migration requires a combination of multiple approaches. The prominent approach in studying natural selection and adaptation, while having access to the single genome of the species is analysis of gene family evolution. We use analysis of gene gains as a way to determine orthogroups potentially evolved in response to adaptation.

Genomics of migration is investigated in great detail for Monarch butterflies. Combination of population, functional and epigenetic studies demonstrate genes involved in orientation, chemoreception, resultation circadian clock mechanism, central complex associated genes are associated in migration. Despite the fact that migratory behavior evolved multiple times in Lepidoptera, some of the above mentioned mechanisms play such a major role, that one may comprae if same of these genes are involved in migration in Vanessa cardui. On the level of gene duplication studies we observed that large 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).

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. Set of orthogroups containing

## Evolution of high adaptability

Looking beyond the question of adaptation to migration in particular we characterized all the gene families which experienced significant changes in evolution of the species. In addition to the above mentioned fat metabolism we observe drastic expansion of gene regulation genes (Figure A a). We see it as a signature of high adaptive capability in expansion of gene regulation genes. In it’s lifestyle Painted Lady needs not only to be capable for long distance fights and orientation, but also be able to adapt to diverse host plants, various habitats, different environmental conditions, temperature.

Local adaptation could be a natural process in this situation, however it may have a lower impact on multigenerational migrant. As environments are drastically changing between generations, passing genes allowing adaptation to the habitat, which was inhabited by parents, may not be beneficial. On the contrary, high plasticity and adaptability of the genome may play a crucial role. Another consequence of this process is maintenance of high heterozygosity (Aurora’s paper).

## Role of recombination in genome evolution

Estimation of the recombination rate is vital for all evolutionary research. Despite the high importance of detailed recombination maps, they are only available for a handful of butterfly species. In several cases linkage maps are used to improve genome assemblies, but analysis of recombination rate is not accessed.   
Genome wide average recombination rate varies from with average value of , which falls into the range of values obtained for other Lepidoptera species. Recombination rate estimated for postman butterflies Heliconius erato equals 6 cM/Mb ([10.1038/sj.hdy.6800619](http://dx.doi.org/10.1038/sj.hdy.6800619)), while closely related species H. melpomene has slightly lower recombination rate: 5.5 cM/Mb (Jiggins et al. 2005). More distant species, silk moth B. mori. average recombination rate of 4.6 cM/Mb (Yasucochi 1998). In other insect species, honey bees, recombination rates significantly exceed these numbers: 19 cM/Mb (doi: 10.1101/gr.5680406). However, it's difficult to compare recombination maps made with different methods and technical aspects should always be considered while comparing recombination rates.

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, in 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).

Unfortunately, we don’t have population data to test this hypothesis.

What factors influence the difference in recombination rate globally and within Lepidoptera? Variation of the recombination rate is observed not only on taxonomic level, but also within the same genome, as it might be a consequence of the difference in number and size of chromosomes. We observed a negative correlation between length of the chromosomes and recombination, also observed in Heliconius (ref)

We went beyond and investigated factors which may shape the recombination landscape in *V.cardui.*

## Recombination landscape in *V.cardui*

What factors shape the recombination landscape in *Vanessa cardui*? As discussed before, 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. In order to address patterns of variation of recombination along the genome one needs to take into account detailed genome architecture: consider positions and spatial distribution of the coding genes and transposable elements (as we don’t have population data). In this study we conclude that the only predictors significantly associated with recombination rate are the position on the chromosome and density of two repeat classes: SINE and non-LTR.

We first addressed correlation between gene content and recombination. It was previously suggested that recombination rate is positively correlated with gene density. Overall, is Vanessa cardui we observe weak positive correlation between gene density and chromosome size, however no correlation is observed between coding sequence and recombination rate (Figure C). For our data, there is no significant correlation between recombination rate and gene density or CDS content. (Figure E)

* + 1. However, we observe the opposite spatial pattern in gene and GC-content distribution throughout the chromosome: it increased towards ends (Figure D)
  1. One of the striking results is that we have strong neg correlation with repeats and chr size too. (Haven’t been addressed before?)

1. Linear model results
   1. 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).
      1. 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.
      2. Interference might explain lower rec rate in the center of the chromosome?
   2. Positive correlation between recombination rate and SINE and (neg) with other non-LTRs. (Figure E, Linear model)
      1. 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.
      2. Discuss ALU (SINE inducing recombination in humans)
      3. Our results are in contrast with this hypothesis: here is a positive correlation between all repeat classes and recombination rate (excluding W). (Figure E)

Discuss differences in repeat annotation

## Gene duplication and genomic architecture

“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”

Above we discussed the process of gene duplication which responses to selective forces and facilitated accumulation of the new genes and patterns of recombination which may not have direct connection to adaptive evolution . Finally, the third power is selfish elements acting on the genomic level and having different influences depending on the class. Integrative analysis allows us to shed light into their interaction.

Chromosome level assembly enables us to look at the process of gene duplications/gains throughout the genome (Figure B c). We observe three patterns: orthogroups spread between several chromosomes (how many?), orthogroups in clusters or within the same chromosome. (Also Figure F)

1. Is pattern caused by the mechanism of duplications?
2. Our findings support the first scenario with mobile elements.
   1. SINEs are negatively correlated with gene gains
   2. LTR, non-LTR, LINE, DNA positive
3. Recombination and gene gains
4. Spatial distribution of repeats (Figure H) demonstrates universal patterns throughout the genome with and exception of W chromosome

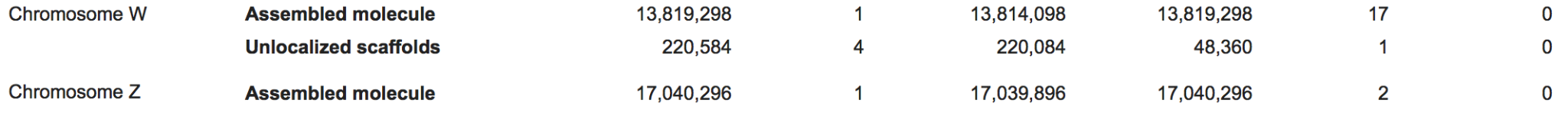
Further study: Timing of the gene duplications (divergence)

W chromosome harbors significantly more gene duplications, due to …

## Evolution of sex-determination system

Classic theory predicts that sex chromosomes originate from a pair of homologous autosomes and recombination between them is suppressed via inversions to resolve sexual conflict. The resulting degradation of the Y chromosome gene content creates the need for dosage compensation in the heterogametic sex.

Transposable elements (TEs) are selfish genetic sequences capable of replicating and inserting themselves throughout the genome. Although often assumed to accumulate following recombination suppression, the insertion of TEs near the sex-determining locus can also act to suppress recombination by creating divergence between sex chromosomes.



In Vanessa cardui W chromosome demonstrates a number of highly specific features: high numbers of TE, and their high length, different distribution of repeat classes compared to other chromosomes (Figure Suppl barplot). On the other hand some signs show degradation of the W chromosome is an intensive ongoing process. First of all, the chromosome is long and has average length compared to average chromosome size. Secondly, many genes that are still present onthe chromosome, it likely that active TE carry and redistribute them within the chromosome. However, it’s difficult to determine if genes on the W chromosome are still functional. The most common orthogroups present on chromosome w are still to be functionally annotated.

Observations from the Z chromosome also support hypothesis that sex chromosomes in Vanessa are relatively young. Z chromosome is the largest chromosome and doesn’t show any significant difference in gene and repeat distribution. 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.

## Conclusions

1. Chromosome level assembly verified by linkage map serve as a resource for genome architecture, organization, population level studies etc
2. Vanessa genome evolved under selection to adapt to migratory lifestyle and due to intreply between different forced effecting it: recombintion, repeats
3. Lots of fat and regulatory genes
4. Recombination landscape in Vanessa cardui
5. W chromosome is very special and opens up opportunities to study chromosome evolution on the large scale (Lepidoptera)
6. 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)](https://www.zotero.org/google-docs/?zhHoqg) . 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)](https://www.zotero.org/google-docs/?WxIIzL). 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.bioin formatics.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)](https://www.zotero.org/google-docs/?broken=yAYxFm). 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). Resulting bam files were sorted with samtools sort (<http://bio-bwa.sourceforge.net>, Li et al. 2009) and 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)](https://www.zotero.org/google-docs/?broken=m77Zsb). 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 minimum mapping quality (-q) 10 and minimum base quality (-Q) 10 [(Li 2011)](https://www.zotero.org/google-docs/?broken=CZ9Y8q). 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)](https://www.zotero.org/google-docs/?broken=9S72PE).

### Construction of linkage map

LepMap3 was used to construct the linkage map [(Rastas 2017)](https://www.zotero.org/google-docs/?broken=3RwSsO). The module ParentCall calls informative parental markers and uses genotype likelihood information from the offspring to impute missing or erroneous parental markers. This module was run with default values, except for zLimit=2 to detect markers segregating as sex chromosomes, and with removal of 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 estimated empirically by testing a range of LODscores (1-30) and set to 12 resulting in 31 linkage groups.

To assign all informative markers to linkage groups JoinSingles was run without informative mask (ie using all markers), and lodLimits 10 deemed optimal, 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 were 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. 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 Darwin Tree of Life assembly we used the software lepAnchor [(Rastas 2020)](https://www.zotero.org/google-docs/?broken=PntZX3). 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](https://doi.org/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.* BLASTwas 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](https://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 *ab initio* 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 14,957. 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 for 13,318 genes. We controlled quality of orthologs alignment using custom filtering and the resulting dataset consisted of 12,098 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)](https://www.zotero.org/google-docs/?Khvjb5). 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)](https://www.zotero.org/google-docs/?XGx5lg).

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)](https://www.zotero.org/google-docs/?OWWxTs). 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.

Likelihoods of all models were compared using Aikaike’s Information Criterion (AIC) calculated as where is the number of parameters and 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)](https://www.zotero.org/google-docs/?DwzBoD) in R version 4.1.0 (R Core Team, [2013](https://onlinelibrary.wiley.com/doi/full/10.1111/mec.15745#mec15745-bib-0071)). We created 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)](https://www.zotero.org/google-docs/?csegFX). 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. We used set of genes associated to biological functions migration in the Monarch butterfly (*Danaus plexippus*), since it’s the only model organism where genomics of migration was investigated in much detail. The gene sets and reference genome were obtained from MonarchBase (<http://monarch.umassmed.edu/resource.html>, accessed 11/06/2021). The nucleotide sequences for the migratory gene set were extracted from reference fasta with a custom script resulting in 588 Monarch genes of interest belonging to five functional categories: genes involved in orientation, chemoreception, resultation circadian clock mechanism, central complex associated genes.

Using the output of OrthoFinder we investigated the presence and absence of genes of interest for migration in *V.cardui* genome annotation (custom scripts). At the next step we used output of BadiRate analysis and identified orthogroups, which experienced extension in *Vanessa cardui.* Genes of *Danaus plexippus* belonging to the same orthogroups were extracted. We then intersected a set of genes belonging to migratory phenotype with genes from orthogroups extended in *Vanessa cardui* and didn’t find overlaps.

## 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. Markers deviating from ascending order were removed with R-script from (<https://github.com/tsackton/linked-selection>) [(Sackton et al. 2014)](https://www.zotero.org/google-docs/?gGzMoB). 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)](https://www.zotero.org/google-docs/?WWMYr3). We transformed the recombination rate per marker to recombination rate in 2 Mb windows across the genome with a custom R script.

### Window-based analysis

We analysed spatial distribution of various genomic features along the *V.cardui* genome using custom scripts. Positions of the repeats (split into classes) were accessed through the RepeatMaker output file and positions of genes were taken from the annotation gff. For the genes belonging to extended orthogroups we integrated output of Badirate (list of the extended orthogroups), OrthoFinder (names of the genes belonging to selected orthogroups) and MAKER (positions of genes of interest).

Genome was split into windows of 2 Mb for the analysis of association with recombination and 100kb for association between genomic features. For each window we calculated density (count/100kb) and occupied length (total number of nucleotides per window) of each of the following: genes, gained genes (genes belonging to extended gene families), repeats in general, SINEs, LINEs, total DNA transposons and specifically Tc1/Mariner, simple repeats)

In-house scripts integrated bash, awk and python/pandas and available on project’s GitHub.

### Correlations between genomic features

Correlation test between different genomic features was performed with cor.test function in R using Spearman’s rank correlation, after testing for normality with Shapiro-Wilk normality test. We used lm function in base R to explore the relationships between ... in a linear model and lmer in the R-package lme4 (ref) for a mixed linear model. Prior to analysis, explanatory variables were scaled and centered.

To analyse the association between gene gains and various 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 values was tested with non-parametric Wilcoxon test implemented in the package rstatix (Alboukadel Kassambara (2021). All the tests were performed two times: once including all windows along the genome, the other excluding the W-chromosome. We used the R-package ggplot2 for visualisation [(Wickham 2009)](https://www.zotero.org/google-docs/?399Wx1).

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

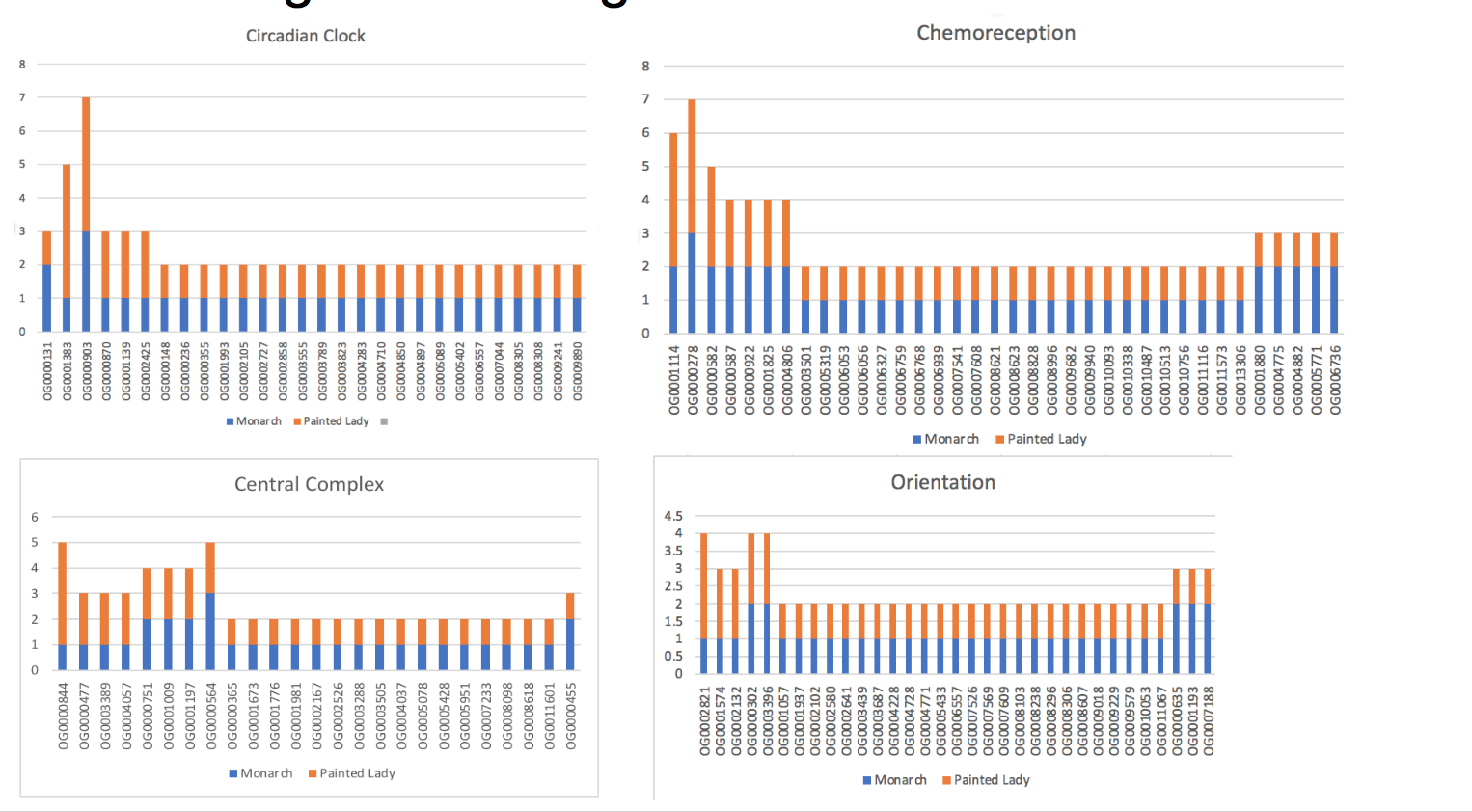
# 

# 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

# 

1. 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)

(Alboukadel Kassambara (2021). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.0. https://CRAN.R-project.org/package=rstatix

(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/](https://www.r-project.org/))

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).](https://www.zotero.org/google-docs/?5BUqZ9)

[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.](https://www.zotero.org/google-docs/?5BUqZ9)

[Aljanabi S. 1997. Universal and rapid salt-extraction of high quality genomic DNA for PCR- based techniques. *Nucleic Acids Res* **25**: 4692–4693.](https://www.zotero.org/google-docs/?5BUqZ9)

[Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**: 238.](https://www.zotero.org/google-docs/?5BUqZ9)

[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.](https://www.zotero.org/google-docs/?5BUqZ9)

[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.](https://www.zotero.org/google-docs/?5BUqZ9)

[Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* **33**: 1635–1638.](https://www.zotero.org/google-docs/?5BUqZ9)

[Librado P, Vieira FG, Rozas J. 2012. BadiRate: estimating family turnover rates by likelihood-based methods. *Bioinformatics* **28**: 279–281.](https://www.zotero.org/google-docs/?5BUqZ9)

[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.](https://www.zotero.org/google-docs/?5BUqZ9)

[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.](https://www.zotero.org/google-docs/?5BUqZ9)

[Wickham H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York https://www.springer.com/gp/book/9780387981413 (Accessed August 24, 2021).](https://www.zotero.org/google-docs/?5BUqZ9)