DR. ROGER VILA (Orcid ID : 0000-0002-2447-4388)
DR. NICLAS BACKSTROM (Orcid ID : 0000-0002-0961-8427)

Article type : Original Article

Lack of gene flow: narrow and dispersed differentiation islands in a triplet of *Leptidea* butterfly species

Venkat Talla[1], Anna Johansson[2], Vlad Dincă[3], Roger Vila[4], Magne Friberg[5], Christer Wiklund[6], Niclas Backström[1,*]

[1]Department of Evolutionary Biology, Evolutionary Biology Centre (EBC), Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden

[2]Department of Medical Biochemistry and Microbiology, Uppsala Biomedical Centre (BMC), Husargatan 3, SE-752 37 Uppsala, Sweden

[3]Department of Ecology and Genetics, PO Box 3000, 90014, University of Oulu, Finland

[4]Institut de Biologia Evolutiva (CSIC-UPF), Pg. Marítim de la Barceloneta 37, 08003 Barcelona, Spain

[5]Department of Biology, Biodiversity Unit, Lund University SE-223 62 Lund, Sweden

[6]Department of Zoology: Division of Ecology, Stockholm University, Svante Arrhenius väg 18B, SE-106 91 Stockholm, Sweden

Emails:

Venkat Talla: venkat.talla[at]ebc.uu.se

Anna Johansson: anna.johansson[at]scilifelab.se

Vlad Dincă: vlad.e.dinca[at]gmail.com

Roger Vila: roger.vila[at]csic.es

Magne Friberg: magne.friberg[at]biol.lu.se

Christer Wiklund: christer.wiklund[at]zoologi.su.se

Niclas Backström: niclas.backstrom[at]ebc.uu.se

*Correspondence: niclas.backstrom[at]ebc.uu.se

Running head

Genomic divergence in cryptic butterflies

Key words

Speciation, wood white butterflies, *Leptidea*, Lepidoptera, cryptic species, genome-scan

Abstract

Genome-scans in recently separated species can inform on molecular mechanisms and evolutionary processes driving divergence. Large-scale polymorphism data from multiple species pairs are also key to investigate the repeatability of divergence – if radiations tend to show parallel responses to similar selection pressures and/or underlying molecular forces. Here we used whole genome re-sequencing data from six wood white (*Leptidea* sp.) butterfly populations, representing three closely related species with karyomorph variation, to infer the species' demographic history and characterize patterns of genomic diversity and

differentiation. The analyses supported previously established species relationships and there was no evidence for post-divergence gene flow. We identified significant intraspecific genetic structure, in particular between karyomorph extremes in the wood white (*L. sinapis*) – a species with a remarkable chromosome number cline across the distribution range. The genomic landscapes of differentiation were erratic and outlier regions were narrow and dispersed. Highly differentiated (F_{ST}) regions generally had low genetic diversity (θ_{π}), but increased absolute divergence (D_{XY}) and excess of rare frequency variants (low Tajima's *D*). A minority of differentiation peaks were shared across species and population comparisons. However, highly differentiated regions contained genes with overrepresented functions related to metabolism, response to stimulus and cellular processes, indicating recurrent directional selection on a specific set of traits in all comparisons. In contrast to the majority of genome-scans in recently diverged lineages, our data suggest that divergence landscapes in *Leptidea* have been shaped by directional selection and genetic drift rather than stable recombination landscapes and/or introgression.

Introduction

The development of DNA sequencing techniques with high yield and comparatively low cost has triggered a massive effort to use population genomics for investigating divergence processes in natural populations (Ravinet *et al.*, 2017; Seehausen *et al.*, 2014; Wolf & Ellegren, 2017). In this pursuit, a wide range of more or less divergent species pairs across the tree of life have been subjected to genome-scan approaches with the main aim to identify differentiation outliers that could indicate barriers to gene flow (Berner & Roesti, 2017; Burri *et al.*, 2015; Delmore *et al.*, 2015; Ellegren *et al.*, 2012; Ferchaud & Hansen, 2016; Harr, 2006; Irwin *et al.*, 2016; Jones *et al.*, 2012; Martin *et al.*, 2013; Poelstra *et al.*, 2014; Renaut *et al.*, 2013; Talla *et al.*, 2017a; Turner *et al.*, 2005; Vijay *et al.*, 2016; Wolf & Ellegren,

2017; Zimmer *et al.*, 2016). One common observation from these genome scans has been that genomic differentiation is highly heterogeneous with pronounced 'differentiation islands' (Berner & Roesti, 2017; Wolf & Ellegren, 2017; Zimmer *et al.*, 2016). This is consistent with the hypothesis that incompatibility genes should be sheltered from interspecific gene flow and therefore stand out as more differentiated than surrounding regions (e.g. Berner & Roesti, 2017; Feder *et al.*, 2013; Nosil *et al.*, 2009; Turner *et al.*, 2005). However, a heterogeneous landscape of genomic differentiation is expected even without gene flow, predominantly as an effect of regional variation in recombination rate (Buerkle, 2017; Burri, 2017; Cruickshank & Hahn, 2014; Ellegren & Wolf, 2017; Jiggins & Martin, 2017; Lohse, 2017; Nachman & Payseur, 2012; Ravinet *et al.*, 2017; Van Doren *et al.*, 2017). Although the exact pin-pointing of genes involved in reproductive isolation likely is out of reach in most natural study systems, a genome-scan approach in recently diverged species can inform on the main molecular mechanisms and evolutionary forces underlying global patterns of genomic divergence during initial stages of the speciation process (Jiggins & Martin, 2017; Seehausen *et al.*, 2014; Wolf & Ellegren, 2017; Zimmer *et al.*, 2016).

First, such studies may reveal candidate regions involved in lineage specific adaptations and inform whether the divergence process involves few genes with large effect, many genes with small effect, or a mix between the two (Jiggins & Martin, 2017). Second, the repeatability of evolutionary trajectories has been a long-standing topic (Conte *et al.*, 2012; Gould, 1990; Meyer *et al.*, 2012). This has been actualized by recent observations of conserved genomic diversity and differentiation across rather distant species pairs (Colosimo *et al.*, 2005; Pereira *et al.*, 2016; Renaut & Dion-Cote, 2016; Renaut *et al.*, 2014; Talla *et al.*, 2017a; Van Doren *et al.*, 2017; Vijay *et al.*, 2016). These studies point towards repeated patterns of diversity and differentiation in independent lineages, likely a consequence of conserved recombination

landscapes in these organisms (Burri, 2017; Ellegren & Wolf, 2017). However, such parallelism could also be a consequence of similar selection pressures in independent radiations, resulting in repeated changes in allele frequency in the same genomic regions. Investigating patterns of genomic differentiation and diversity across species complexes with different divergence times will aid in understanding the prevalence of both forms of parallelism in divergence processes.

The three butterflies wood white (*Leptidea sinapis*), Real's wood white (*Leptidea reali*) and cryptic wood white (*Leptidea juvernica*) represent one of the most striking examples of cryptic species in Eurasian butterflies (Dincă *et al.*, 2011; Dincă *et al.*, 2013). The estimated divergence times between species range from 1-2 million years (*L. sinapis* vs. *L. reali*) to 2.5-3.5 million years (*L. juvernica* vs. the two other species) (Talla *et al.*, 2017b). The three species are indistinguishable in the field, only partially discernible by genitalia (*L. sinapis* from the other two) and karyotype (*L. reali* from *L. juvernica*), but can be reliably identified by DNA analysis. Overall, the species in the complex are distributed over a vast geographic range (Figure 1), covering a wide variety of habitats. They exhibit complex differences in life history characteristics, pupal and genital morphology, mating behaviour, pheromone profiles, habitat preference, karyotype set-up and genome size (Dincă *et al.*, 2011; Dincă *et al.*, 2013; Freese & Fiedler, 2002; Friberg, 2007; Friberg *et al.*, 2008a; Friberg *et al.*, 2008b; Friberg & Wiklund, 2010; Lorkovic, 1993; Lukhtanov *et al.*, 2011; Šíchová *et al.*, 2015; Talla *et al.*, 2017b; Wiklund, 1977a, 1977b). Moreover, both *L. sinapis* and *L. juvernica* are further subdivided into multiple geographically distinct ecotypes with complex habitat utilization preferences and life-history characteristics, potentially reflecting variation in environmental conditions across the distribution ranges (Friberg *et al.*, 2013; Friberg *et al.*, 2008a; Friberg & Wiklund, 2009, 2010). *Leptidea sinapis* also shows striking intraspecific chromosome

number variation present in a cline with a gradual increase from 2n ≈ 56 - 58 in the northern and eastern parts (Scandinavia – Kazakhstan) to 2n ≈ 106 - 108 in the southwestern parts (Iberia) of the distribution range (Dincă *et al.*, 2011; Dincă *et al.*, 2013; Šíchová *et al.*, 2016; Šíchová *et al.*, 2015). The exact mechanism behind this karyotype variation is unknown, but both visual inspection of metaphase spreads (Lukhtanov *et al.*, 2011) and genome size analysis (Talla *et al.*, 2017b) strongly suggest that this is a result of a large number of chromosome fissions/fusions rather than whole or partial genome duplications. *Leptidea sinapis* individuals belonging to different karyotypes can reproduce when brought together, although offspring fitness is considerably reduced for F2 and later generations if the most distant karyotypes are crossed (Lukhtanov *et al.*, 2018). This has raised the prospect that, although still part of the same species, different local populations have accumulated intrinsic reproductive barriers, potentially as a result of a combination of local adaptation and chromosome rearrangements.

Here we applied a population genomics approach to quantify patterns of genetic diversity and differentiation across the triplet of wood white cryptic species. The primary aims of the study were to, i) understand the demographic history of the three species, ii) investigate if gene flow, underlying molecular mechanisms or adaptive processes have been the main drivers of divergence, and iii) estimate genome wide statistics of diversity and differentiation to identify regions underlying potential lineage-specific adaptations.

Materials and methods

Sampling and sequencing

A total of 60 specimens representing *L. sinapis*, *L. reali* and *L. juvernica* were sampled (Figure 1). These included 10 male individuals from each of six populations (*L. sinapis* Sweden (LsSwe), *L. sinapis* Spain (LsSpa), *L. sinapis* Kazakhstan (LsKaz), *L. reali* Spain (LrSpa), *L. juvernica* Kazakhstan (LjKaz), and *L. juvernica* Ireland (LjIre)), representing both allopatric and sympatric species pairs as well as the most extreme karyotypes within *L. sinapis* (Figure 1). Species identity was determined based on genital morphology (Dincă *et al.*, 2011; Lukhtanov *et al.*, 2011). Individually barcoded 380 bp paired-end libraries were prepared, multiplexed and sequenced using Illumina HiSeq technology (Illumina, Inc., San Diego, USA). A reference genome was assembled from a combination of paired-end and mate-pair libraries from a five-generation full-sib inbred male Swedish *L. sinapis*. The reference genome was 643 Mb and the average genome sequence coverage across individuals was 12X. For details regarding the library preparations, sequencing and genome assembly, see Talla *et al.* (2017b).

Read mapping and polymorphism scoring

The reads obtained were trimmed for adapter sequences and low quality bases using Cutadapt v. 1.14 (Martin, 2011). Trimmed reads were mapped to the reference genome assembly with BWA v. 0.7.12 (Li & Durbin, 2010) using the algorithm "mem", and resulting .sam files were pre-processed in SAMtools v. 1.2 (Li *et al.*, 2009). Local insertion/deletion (indel) realignment and duplicate marking was done using GATK v. 3.2.2 (McKenna *et al.*, 2010) and final mapping qualities were assessed using BamQC (Andrews, 2016). The mapping success across samples was very high; 97.9 ± 0.6% in *L. sinapis*, 97.5 ± 0.45% in *L. juvernica* and 98.56 ± 0.08% in *L. reali*. The average genome coverage was 12.9X for *L.*

sinapis, 12.7X for *L. juvernica* and 10.9X for *L. reali* (Supplementary Table 1). Alignments were improved by removing duplicate reads using MarkDuplicates in Picard tools v. 1.127 (http://broadinstitute.github.io/picard/), and realigning around problematic regions using RealignerTargetCreator / IndelRealigner in GATK v. 3.4.46 (McKenna *et al.*, 2010). With the aim of finding a 'golden set' of polymorphisms that could be used for further recalibration of the data, variants were called using both HaplotypeCaller (McKenna *et al.*, 2010), FreeBayes v. 0.9.10 (Garrison & Marth, 2012) and SAMtools v. 1.2 (Li *et al.*, 2009). All singleton variants were removed, and only variants homozygous in at least one individual and scored by at least two methods were kept. This set contained 14,367,566 variants and was used as input for the GATK BaseRecalibrator / PrintReads before individual genotyping was performed with HaplotypeCaller, where individuals were genotyped together with the GenotypeGVCFs in GATK v. 3.4.46 (McKenna *et al.*, 2010). SNPs with the highest quality (10%) were then used for variant quality score recalibration (VQSR) filtering using the VariantRecalibrator in GATK v. 3.4.46 (McKenna *et al.*, 2010). Only variants that passed the VQSR filtering were used for down-stream analysis. The work-flow for variant calling is presented in Supplementary Figure 1.

Analysis of population structure

The mitochondrial genomes were assembled for each individual separately and aligned using MAFFT v. 7 (Katoh & Standley, 2013). A mitochondrial DNA (mtDNA) phylogeny was inferred using the Neighbor-Joining method (Saitou & Nei, 1987) with distances estimated using Maximum Composite Likelihood (Tamura *et al.*, 2004) and among sites rate variance modelled using a gamma distribution ($\alpha = 1$) as implemented in MEGA7 (Kumar *et al.*, 2016). Missing positions were removed for pairwise comparisons.

To further characterize potential genetic structure among populations and species we used both a maximum likelihood clustering approach that estimates ancestries based on specified number of clusters as implemented in ADMIXTURE (Alexander *et al.*, 2009), and principal component analysis (PCA) of the nuclear genetic variance across individuals. To minimize any potential bias due to coverage variation we only used SNPs covered in all individuals in each population for these analyses. ADMIXTURE was run with default settings and the PCA was conducted using the R/Bioconductor package SNPRelate (Zheng *et al.*, 2012) after filtering out sites with linkage disequilibrium (r^2) > 0.2. Both ADMIXTURE and PCA analysis were initially conducted with all 60 individuals from the six populations, and separate intraspecific analyses were subsequently conducted for the 30 *L. sinapis* individuals and the 20 *L. juvernica* individuals, respectively. The optimal number of clusters in the ADMIXTURE analysis was specified based on the lowest cross validation error rates (Alexander *et al.*, 2009; Evanno *et al.*, 2005).

Introgression analysis

The ABBA/BABA approach, as implemented in ANGSD (Korneliussen *et al.*, 2014), was used to identify potential introgression between the sympatric population pairs LsSpa – LrSpa and LsKaz – LjKaz. To polarize the ABBA/BABA informative polymorphisms and calculate the Patterson's *D*-statistic [(nABBA-nBABA)/(nABBA+nBABA)], we used LjKaz and *Pieris rapae* as outgroups for each test, respectively (Supplementary Figure 2). *Pieris rapae* belongs to a different Pieridae subfamily (Pierinae) than *Leptidea* (Dismorphiinae). The genomic reads used to generate the *P. rapae* reference were obtained from the Sequence Read Archive (SRA; accession number SRR4339879). The *P. rapae* reads were mapped to the *L. sinapis* reference genome using Stampy (Lunter & Goodson, 2011) and the output .bam files were converted to .fasta format using ANGSD (Korneliussen *et al.*, 2014). For

both analyses, only the sample with the highest genome wide average coverage in each population was used. ABBA and BABA sites were counted in 10 kb windows across the genome in both the tests, and genomic blocks were then bootstrapped using an R-script (jackKnife.R) provided by ANGSD to obtain the mean and variance of the *D*-statistic (Korneliussen *et al.*, 2014).

Analysis of past population size changes

To trace potential historical fluctuations in population size, a pairwise sequentially Markovian coalescent (PSMC) model analysis was applied. The method extracts information about temporal variation in coalescence times estimated from the genome-wide distribution of polymorphisms in a single diploid genome (Li & Durbin, 2011). Parameter settings were set according to recommendations (Li & Durbin, 2011) for a range of recombination / mutation (ρ / θ) ratios from 0.1 – 5.0. Average generation time was set to one year and the mutation rate to $2.9*10^{-9}$ per site per generation (Keightley *et al.*, 2014).

Population genomic analysis

We inferred levels of mtDNA nucleotide diversity by calculating the average sequence differences over all sequence pairs (π) in each population, omitting missing data in pairwise comparisons, using the Maximum Composite Likelihood method (Tamura *et al.*, 2004) as implemented in MEGA7 (Kumar *et al.*, 2016). In addition, mtDNA divergence across population pairs (average number of differences between all inter-population pairs of sequences, missing data omitted) was estimated using the Maximum Composite Likelihood method (Tamura *et al.*, 2004) as implemented in MEGA7 (Kumar *et al.*, 2016).

For the nuclear genomic data, a set of different population genetic summary statistics were calculated based on SNPs (including invariant sites). To minimize effects of potential false genotype calls, only sites covered in all individuals from a population were used to calculate nucleotide diversity (θ_π) and Tajima's D (T_D) (Tajima, 1989) within each population. Similarly, for pairwise comparisons between species and populations (F_{ST} and D_{XY}), only sites covered in all 10 individuals in all included populations were used. T_D, θ_π and F_{ST} (Weir & Cockerham, 1984) were estimated using vcftools (Danecek *et al.*, 2011) and D_{XY} was estimated using an in-house developed python script (see data accessibility paragraph). Rates of fixed, shared and private polymorphisms for pairwise species- and population comparisons were calculated from the allele frequency estimates generated in vcftools. To identify regions in the genome linked to potential lineage specific adaptations we combined information about differentiation (F_{ST}) and absolute divergence (D_{XY}). Relative F_{ST} (F_{ST}^{Z}: (window F_{ST} - genome-wide F_{ST}) / standard deviation of genome-wide F_{ST}) scores were estimated for each 10 kb window and the top 1% were considered differentiation outliers. If a F_{ST}^{Z} outlier window also had a higher than genomic average D_{XY} value it was considered a candidate window for containing genes under divergent selection. Note that lineage specific directional selection is not expected to generate elevated D_{XY}, this is just a means to avoid including loci with reduced diversity in the ancestral lineage (before divergence). Binomial sampling, as implemented in R, was used to assess potential overrepresentation of overlapping outlier windows in the F_{ST}^{Z} and T_D distributions across pair-wise population and species comparisons.

All population genetic summary statistics were estimated in non-overlapping 10 kb windows across all scaffolds. Windows with less than 10% of the bases covered in any individual included in the analysis were removed from downstream analysis. Species level summary

statistics were estimated by including all individuals from a species in the analysis, implementing the filtering threshold for sites of ≥ 7X coverage in ≥ 7 individuals from each population. For visualization purposes, all scaffolds in the *L. sinapis* genome assembly were aligned to the genome assembly of *Heliconius melpomene* to anchor the scaffolds on chromosomes. The assignment to chromosomes was based on the best (lowest e-value) BLAST hit of a *L. sinapis* scaffold to a *H. melpomene* chromosome. Although this procedure inevitably results in some erroneous assignments due to chromosome rearrangements between *L. sinapis* and *H. melpomene*, the lepidopteran karyotype is generally very stable (Ahola *et al.*, 2014; The Heliconius Genome Consortium, 2012) and the quantitative results and the interpretations in this study should only be marginally affected.

Ontology enrichment analysis

The top 1% of the 10 kb windows with highest relative differentiation (F_{ST}^{Z}) and higher than genomic average D_{XY} were used for an outlier gene enrichment analysis. Genes within outlier windows were identified by BLAST to the entire gene set of *Drosophila melanogaster* (obtained from FlyBase; http://flybase.org/), and a gene-ontology enrichment analysis for genes in outlier regions was done using the PANTHER data base (pantherdb.org; Mi *et al.*, 2016).

Results

Population structure and speciation history

To provide a clear picture of species and population relationships and potential variation in past demographic events that could influence patterns of genome differentiation, the mtDNA and genome-wide SNP data were initially used to assess global relationships, genetic clustering, historical variation in N_e and potential gene flow between sympatric species pairs.

The mtDNA phylogeny recovered each of the three species as monophyletic, *L. juvernica* being the outgroup to the sister species *L. sinapis* and *L. reali*. Populations of LsSpa, LrSpa, LjIre and LjKaz were reciprocally monophyletic, while LsSwe and LsKaz showed a paraphyletic pattern (Figure 2). The analyses of genetic clustering with ADMIXTURE and PCA using genomic data showed consistent results with the mtDNA tree; all species clustered in distinct groups and the cross-entropy error rates in ADMIXTURE were lowest for K-values 3 – 5 (Figure 2). At the population level, again the PCA and the ADMIXTURE analysis revealed intraspecific structure within *L. sinapis* and *L. juvernica* with LsSpa being distinct from LsSwe and LsKaz, and LjIre being distinct from LjKaz (Figure 2, Supplementary Figure 3). When all samples were analysed jointly, there was no indication of structure between LsSwe and LsKaz but the two populations were separated in both the PCA and ADMIXTURE analyses when only the three *L. sinapis* populations were included in the analysis (Supplementary Figure 4). The ABBA/BABA tests failed to identify any traces of introgression between species in sympatric regions; LsSpa – LrSpa D = -0.012 and LsKaz – LjKaz D = -0.021. This indicates that all three species are genetically distinct, with no evidence for post divergence gene flow, and that there is considerable genetic structure between LjKaz and LjIre and between karyotype extremes within *L. sinapis*. The analysis of past demographic changes in population size revealed considerable differences across species and populations. Both *L. reali* and *L. juvernica* showed marginal fluctuations in N_e over time with some evidence for recent declines in LjIre and LrSpa. All *L. sinapis* populations showed a marked increase in N_e over the last 10s of thousands of generations with the most dramatic expansion in LsSwe (Figure 2, Supplementary Figure 5).

Global genetic diversity and allele frequency distributions in species/populations

To obtain a detailed picture of global levels of genetic diversity and allele frequency distributions, both at the species level and for all populations separately, θ_π and T_D were estimated for both the mtDNA genomes and in non-overlapping 10 kb windows across all nuclear scaffolds. The mtDNA diversity varied extensively across populations, being lowest in LjIre and LrSpa and highest in LjKaz (Supplementary Table 2). We identified a total of ~ 8.51 million nuclear SNPs in *L. sinapis*, ~ 4.20 million in *L. reali* and ~ 3.35 million in *L. juvernica* (Table 1). In line with this, θ_π varied across species being highest in *L. sinapis*, followed by *L. reali* and *L. juvernica* (Table 1, Figure 3). When populations were analysed separately, θ_π and T_D followed the expectations from the inference of historical changes in N_e, with comparatively low θ_π in LjIre, LjKaz and LrSpa while all *L. sinapis* populations showed higher θ_π (Figure 3) and a lower T_D (Table 1, Figure 4, Supplementary Figure 6). As expected from a smaller N_e as compared to autosomes, the Z-chromosome had a lower θ_π in all populations (Table 1); the ratio of Z-chromosome to autosomal θ_π ranged from ~ 0.73 (both LsKaz and LrSpa) to ~ 0.87 (both LjKaz and LjIre).

Global levels of genetic differentiation and absolute divergence across species/populations

The number of fixed differences supported the previously established phylogenetic relationship between species (Supplementary Table 3). As expected, given the substantially lower genetic diversity in *L. reali* and *L. juvernica* as compared to *L. sinapis*, an overall higher F_{ST} was observed between *L. reali* and *L. juvernica* (0.28 ± 0.081) than between *L. sinapis* and any of the other two species (0.15 ± 0.055 vs. *L. reali*; 0.16 ± 0.044 vs. *L. juvernica*) (Figure 4, Figure 5, Supplementary Figure 6). In contrast, although D_{XY} varied considerably across genomic regions in all species and population comparisons, similar levels of average D_{XY} were observed across the three species comparisons (Table 2, Figure 4,

Supplementary Figure 6, Supplementary Figure 7). In line with differences in effective population size across chromosome classes, F_{ST} was significantly lower on the autosomes as compared to the Z-chromosome (Table 2, Figure 5; Mann-Whitney U-tests, p-values: $8.4*10^{-296}$ – $8.3*10^{-44}$).

Regional variation in genetic diversity, differentiation and divergence

Nucleotide diversity was heterogeneous across the genome, both at the species (Figure 4) and population level (Supplementary Figure 7), and there were significant positive correlations in regional θ_π between species and population pairs (Pearson's r-values > 0, p-values < 0.001; Supplementary Figure 8). In agreement with varying levels of θ_π, F_{ST} and D_{XY} also varied extensively across the genome in pairwise species- and population comparisons (Figure 4, Supplementary Figure 7). There was a significant positive correlation between θ_π and D_{XY} and a significant negative correlation between θ_π and F_{ST} in all comparisons (p-values < 0.001; Figure 6, Supplementary Figure 9). The relative differentiation (F_{ST}^{Z}) outliers had on average lower θ_π, but higher absolute D_{XY}, than other regions, in all species and population pair comparisons, but the difference in D_{XY} was not significant for the interspecific analysis involving *L. juvernica* (Table 3, Figure 6, Supplementary Table 4, Supplementary Figure 9). In all comparisons, F_{ST}^{Z} outlier windows with higher than average absolute D_{XY} also had significant excess of rare alleles indicated by a low T_D (p-values < $2.2*10^{-16}$; Supplementary Table 5). These patterns indicate that post-divergence lineage specific processes rather than linked selection in the ancestral lineage have been the main drivers of genome differentiation in this species complex.

There was limited overlap in outlier regions between comparisons (Supplementary Table 6). Specifically, only 17 outlier windows were observed across all species comparisons. Within *L. sinapis*, LsSpa had a large number of lineage specific outlier windows (n = 204) and there were only two outlier windows that were identified in all three population comparisons (Supplementary Table 6). There was also limited clustering of outlier windows in all comparisons (Supplementary Table 7). Although the numbers of adjacent outlier windows were higher than expected based on random sampling, the maximum ranged between four and six adjacent windows in a single comparison (Supplementary Table 7). Hence, most outlier windows were narrow and scattered across the genome and extremely few outlier regions were detected across all species or population comparisons.

Genome differentiation patterns

As already indicated in the population structure and phylogenetic analyses, we observed significant genetic differentiation between some population pairs within the same species. LsSwe and LsKaz, which have similar karyotype set-ups (2n = 56 - 58) but are geographically far apart, showed a low level of genetic differentiation (F_{ST} = 0.019 ± 0.030). Comparisons between LsSwe or LsKaz on the one hand and LsSpa (2n = 106 - 108) on the other, revealed considerably higher levels of genetic differentiation (F_{ST} = 0.13 - 0.14; Figure 5). The two *L. juvernica* populations, LjIre and LjKaz, were moderately genetically differentiated (F_{ST} = 0.09 ± 0.056; Figure 5, Supplementary Figure 7). This allowed us to use independent intraspecific population pairs to investigate if the same regions have been involved in driving genomic divergence between lineages. Although the correlations of genetic differentiation in independent comparisons were significant, as expected given the extremely large number of data points, there were no strong associations (Pearson's *r* varied

from 0.058 to 0.075) between particular genomic regions and the level of genetic differentiation in independent species or population comparisons (Supplementary Figure 10).

Outlier analyses to detect genomic regions under selection

The 'outlier' windows, i.e. the windows in the top 1% of the distribution of F_{ST}^{Z} which also had a higher than average D_{XY}, were screened to identify potential target genes for lineage specific adaptations. In the species pair comparisons, we identified 20 genes in the scan using *L. sinapis* vs. *L. reali*, 18 genes in *L. reali* vs. *L. juvernica* and 25 genes in *L. sinapis* vs. *L. juvernica* (Supplementary Table 8A). The gene ontology enrichment analysis revealed that cellular and metabolic processes, as well as response to stimulus (GO:0009987, GO:0008152, GO:0050896) were most significantly overrepresented in all three comparisons (Supplementary Table 9A). In the comparisons between intraspecific population pairs we identified 14 genes in LsKaz vs. LsSwe, 29 genes in LsSpa vs. LsSwe, 22 genes in LsKaz vs. LsSpa, and 9 genes in LjIre vs. LjKaz (Supplementary Table 8B). Consistent with the species level comparisons, cellular and metabolic processes came out as the most overrepresented gene ontology terms in all of the population-level comparisons and response to stimulus in two out of three (Supplementary Table 9B).

Discussion

Speciation and demographic history

Our data supported previously established taxonomic groups and their relationships, both at the species level and for the intraspecific structure identified within the cryptic wood white, *L. juvernica* (Dincă *et al.*, 2011; Dincă *et al.*, 2013; Talla *et al.*, 2017b) and the common wood white, *L. sinapis* (Talla *et al.*, 2017b). Within *L. sinapis*, there was limited genetic differentiation between the Swedish (2n = 57, 58) and the Kazakhstan populations (2n = 56-

64) despite being separated by a geographic distance > 4,000 km, while both of these populations were considerably differentiated from the Spanish population (2n = 106, 108). This suggests that the chromosome number cline observed across Eurasia for this species (Dincă *et al.*, 2011; Dincă *et al.*, 2013; Šíchová *et al.*, 2016; Šíchová *et al.*, 2015) likely has evolved as a consequence of secondary contact after separation in discrete glacial refugia. A hypothetical but straightforward scenario could be that the ancestral population – probably harbouring a karyotype similar to the ancestral Lepidoptera type (2n approximately = 60; Ahola *et al.*, 2014) as seen in the sister species, *L. reali* (2n = 52-54) – was widespread across Eurasia and that glacial intervals forced individuals to refugia in both the Iberian Peninsula and south-central Asia. Due to low effective population size during glaciation periods, the Spanish population experienced fixation of multiple chromosomal fissions, leading to a larger number of chromosomes. Recolonization then occurred, with a more dramatic expansion from the refugium in the east all the way to north-central Europe (Scandinavia), so that populations with discrete karyotypes met in an elongated contact zone from south-central to central Europe. Hybrids with intermediate karyotypes which are still fertile (Lukhtanov *et al.*, 2018), then recurrently backcrossed into parental populations on each side of the contact zone, generating the chromosome number cline that currently stretches from south-west towards north and east (Dincă *et al.*, 2011; Dincă *et al.*, 2013; Šíchová *et al.*, 2015). Both the analyses of demographic changes and the allele frequency distributions support such a scenario, with a more dramatic expansion in Swedish and Kazakhstan *L. sinapis*.

Discovery of a hybrid specimen between *L. sinapis* and *L. reali* in the sympatric region in Catalonia, Spain (Vlad Dincă, unpublished observation) and indications of potential hybridization events between *L. juvernica* and *L. sinapis* in south-eastern Europe (Solovyev *et al.*, 2015; Verovnik & Glogovcan, 2007) show that hybridization can occur. However, the

visual inspection of genetic clustering (PCA), the admixture analysis and the ABBA/BABA approach all failed to identify any traces of post-divergence gene flow between any species pair. This indicates that reproductive barriers are virtually complete between the species, despite comparatively short divergence times (Talla *et al.*, 2017b) and that the few hybrids that may exist are largely or completely infertile. The three species differ in several aspects that could be directly linked to reproductive isolation; *L. juvernica* males have a distinct display behaviour during courtship (Dincă *et al.*, 2013; Friberg *et al.*, 2008b), *L. sinapis* shows considerable size and shape differences in genital morphology as compared to the other species (Dincă *et al.*, 2011), *L. sinapis* and *L. juvernica* show geographically variable habitat preferences (Friberg *et al.*, 2013; Friberg *et al.*, 2008a; Friberg & Wiklund, 2009), and all species have distinct karyotype structures (Dincă *et al.*, 2011; Lukhtanov *et al.*, 2011). It is tempting to speculate that changes in karyotype should provide the key factor for isolation between the three species, especially since fissions and/or fusions also have involved the sex-chromosomes (Šíchová *et al.*, 2016). Chromosome rearrangements should generally lead to segregation distortion problems during meiosis in F1 hybrids, resulting in reduced hybrid fertility. However, even the most extreme karyotype variants in *L. sinapis* (2n = 56 and 108, respectively) can be crossed and the hybrids can produce fertile offspring via a mechanism of inverted meiosis (Lukhtanov *et al.*, 2018). Consequently, it is not obvious that the chromosome number variation across *Leptidea* species – which is actually less pronounced than between *L. sinapis* population extremes – has generated a complete barrier, although it seems to have contributed to it. All three species show strong pre-mating reproductive isolation that is maintained by female acceptance of conspecific males exclusively (Dincă *et al.*, 2013; Friberg *et al.*, 2008b). If reproductive isolation evolved in allopatry prior to secondary contact, then the power to identify barrier loci will be limited, because the genome divergence landscape will not be affected by regional homogenizing effects of introgression

(Cruickshank & Hahn, 2014; Irwin *et al.*, 2016; Ravinet *et al.*, 2017; Wolf & Ellegren, 2017). However, characterization of genome-wide patterns of population genetic summary statistics may still lead to a more comprehensive picture of the relative importance of different factors driving genomic divergence between incipient species.

Global patterns of genetic variation

The global levels of genetic diversity in all species were in the range of $2\text{-}4*10^{-3}$, indicating that long term effective population sizes have been small compared to many other butterfly species (e.g. Ahola *et al.*, 2014; Cong *et al.*, 2016; Martin *et al.*, 2016). The diversity levels of mtDNA were reduced as compared to nuclear genome levels, as expected since the mitochondria are uniparentally inherited and lack recombination.

As a consequence of the considerable variation in global levels of genetic diversity across species and populations, the genome wide estimates of genetic differentiation, absolute divergence and the number of fixed differences also varied across species and population comparisons. In essence, a reduced genetic diversity in one or both populations used for estimating differentiation, inherently leads to elevated F_{ST} levels, since F_{ST} is calculated as the proportion of genetic variation explained by interpopulation variation scaled by the average level of variation in each respective population (Wright, 1965). In our data, this leads to a higher overall level of genetic differentiation between for example *L. reali* and *L. juvernica,* than between *L. sinapis* and *L. juvernica,* although the divergence time between these species pairs is the same – i.e. *L. juvernica* is an outgroup to *L. sinapis* and *L. reali* (Dincă *et al.*, 2011). The difference in diversity across species, reflects differences in long-term N_e, which affects the fixation rate. The number of fixed differences was smaller in the comparison involving *L. sinapis – L. juvernica* than in the comparison of *L. juvernica – L.*

reali. This is most likely a consequence of a larger N_e, and a lower rate of allele loss in *L. sinapis*.

We observed a significantly reduced genetic diversity on the Z-chromosome as compared to the autosomes in all species and populations. This was also reflected in a significantly higher level of genetic differentiation on the Z-chromosome. This is expected based on the lower effective population size of the Z-chromosome which, in Lepidoptera, is found in one copy in females (ZW) and two copies in males (ZZ). At equal sex ratios, the heterogametic state in females leads to a 25% reduction in N_e for the Z-chromosome as compared to autosomes. However, there is reason to believe that males have higher reproductive variance than females in *Leptidea* species (Wiklund & Solbreck, 1982), leading to an even more pronounced reduction in N_e for the Z, since fewer males than females contribute genetic variation passed on to the next generation. This reduction in N_e should be counteracted by the expected higher rate of recombination on the Z-chromosome compared to the autosomes. The reason for this is that female achiasmy (lack of recombination in female meiosis) is ubiquitous in Lepidoptera (Marec, 1996). Since the Z-chromosome spends 2/3 of the time, from an evolutionary perspective, in the male germline, it will be affected by the male recombination rate more than the autosomes, which spend ½ of the time in the male and ½ in the female germline. Interestingly, the observed ratio of Z/autosome genetic diversity was above 73% in all populations. This indicates that the reduction in N_e as a joint effect of lower chromosome copy numbers in the population and a female biased operational sex ratio, is counteracted, and actually overcompensated for, by a higher overall cross-over rate of the Z-chromosome as compared to the autosomes.

Regional variation in diversity, differentiation and divergence

To investigate the mechanisms driving the differentiation processes, we conducted a genome-scan analysis where population genetic summary statistics were calculated in 10 kb non-overlapping windows. Since F_{ST} is heavily influenced by the θ_π within populations, we also used D_{XY} as a measure of cross species and cross population divergence. Our results are in contrast to many previous genome scan analyses of closely related species (Burri, 2017; Ellegren & Wolf, 2017; Irwin *et al.*, 2016; Ravinet *et al.*, 2017). First, we did not observe pronounced differentiation islands in any species or population comparison. The windows that had high relative genetic differentiation were only marginally clustered together and never spanned large (Mb scale) genomic regions as has commonly been observed in other studies (Wolf & Ellegren, 2017). These windows also had lower genetic diversity than other regions. This pattern is expected as a consequence of either i) increased drift due to low recombination (Hill & Robertson, 1966) or ii) divergent selection in either or both lineages compared, or both forces combined. Second, in contrast to an effect of a conserved underlying recombination landscape and linked selection in the ancestral population prior to species/population divergence, the outlier windows also in general had higher than average absolute divergence and an excess of rare frequency variants. A recombination landscape with considerable regional rate variation, that is stable across lineages, leads to a regional reduction in N_e in orthologous (low recombination rate) regions and higher effect of genetic drift in the same regions in different lineages (Hill & Robertson, 1966). Typically, this results in elevated genetic differentiation but reduced absolute divergence as compared to the genomic average (Burri, 2017; Cruickshank & Hahn, 2014; Ravinet *et al.*, 2017). The recombination landscape is therefore likely rather even across *Leptidea* chromosomes, similar to what recently has been found in bumblebees outside of centromeres (Kawakami *et al.*, 2019). Alternatively, if there is considerable regional variation in recombination rate, the high

and low recombination regions may be highly ephemeral. Both of these options would be minimizing the effect of linked selection reducing diversity (Hill & Robertson, 1966), in particular regions with low recombination rate in all species. Discriminating between these two scenarios would require detailed recombination maps for each respective population, something that is currently lacking in this system. Third, the genetic differentiation outliers were often specific to each respective species and population comparison and not shared across all comparisons. This is different from the situation in many other systems, where genome scan approaches have identified more or less the same set of outlier regions even when independent species comparisons have been made across very divergent families (Van Doren *et al.*, 2017; Vijay *et al.*, 2016). Taken together, the divergence landscapes in *Leptidea* seem to be shaped by lineage-specific selection and genetic drift, generating narrow and dispersed differentiation peaks. Hence, virtually complete reproductive barriers were likely established already before secondary contact. This is in contrast to the vast majority of recent genome-scans in many other lineages, where conserved recombination landscapes and/or recurrent introgression have been key factors shaping landscapes of genomic divergence (Berner & Roesti, 2017; Burri *et al.*, 2015; Delmore *et al.*, 2015; Ellegren *et al.*, 2012; Ferchaud & Hansen, 2016; Harr, 2006; Irwin *et al.*, 2016; Jones *et al.*, 2012; Martin *et al.*, 2013; Poelstra *et al.*, 2014; Ravinet *et al.*, 2017; Renaut *et al.*, 2013; Talla *et al.*, 2017a; Turner *et al.*, 2005; Vijay *et al.*, 2016; Wolf & Ellegren, 2017; Zimmer *et al.*, 2016).

Functions of genes in differentiation outlier regions

In a first attempt to characterize the gene functions that might be involved in lineage specific adaptive processes, we selected the most differentiated regions in each respective species and intraspecific population comparison and i) investigated if the same regions showed elevated differentiation in multiple comparisons and ii) characterized the functions of genes located in

differentiation outliers in all species and intraspecific population comparisons. We found that outlier regions often were unique to specific comparisons and the level of differentiation was only marginally correlated in independent pair-wise population analyses. This indicates that the parts of the genome that diverge at the highest rate are distinctive in each lineage. From a positional point of view, there is hence no parallelism, indicating that divergence involves different target loci in different species and populations. However, there is considerable overlap in the ontology terms associated with outlier regions in all species and population comparisons. The three main overlapping terms are cellular process, metabolic process and response to stimulus, which are among the top ontology terms in all analyses. Key differences between species (and to some extent populations) involve habitat preference, host plant utilization and mate recognition. Response to stimuli (light, chemical cues, temperature) and metabolic rate and efficiency (host plant usage, temperature) are obviously important processes for optimization of habitat usage and can be directly linked to these key differences across the three wood whites.

Acknowledgements

This work was supported by a junior research grant from the Swedish Research Council (VR) to NB. The authors acknowledge support from the National Genomics Infrastructure in Stockholm and Uppsala funded by the Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council, and SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. RV was supported by project CGL2016-76322-P (AEI/FEDER, UE). We thank the associate editor and three anonymous reviewers for detailed and constructive comments that improved the quality of the manuscript.

References

Ahola, V., Lehtonen, R., Somervuo, P., Salmela, L., Koskinen, P., Rastas, P., . . . Hanski, I. (2014). The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nature Communications, 5*, e4737. doi:10.1038/ncomms5737

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research, 19*, 1655–1664.

Andrews, S. (2016). BamQC. Retrieved from https://github.com/s-andrews/BamQC

Berner, D., & Roesti, M. (2017). Genomics of adaptive divergence with chromosome-scale heterogeneity in crossover rate. *Molecular Ecology, 26*, 6351-6369. doi:10.1111/mec.14373

Buerkle, C. A. (2017). Inconvenient truths in population and speciation genetics point towards a future beyond allele frequencies. *Journal of Evolutionary Biology, 30*, 1498-1500. doi:10.1111/jeb.13106

Burri, R. (2017). Linked selection, demography and the evolution of correlated genomic landscapes in birds and beyond. *Molecular Ecology, 26*, 3853-3856. doi:10.1111/mec.14167

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., . . . Ellegren, H. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Research, 25*, 1656-1665. doi:10.1101/gr.196485.115

Colosimo, P. F., Hosemann, K. E., Balabhadra, S., Villarreal, G., Jr., Dickson, M., Grimwood, J., . . . Kingsley, D. M. (2005). Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science, 307*, 1928-1933.

Cong, Q., Shen, J., Warren, A. D., Borek, D., Otwinowski, Z., & Grishin, N. V. (2016). Speciation in cloudless sulphurs gleaned from complete genomes. *Genome Biology and Evolution, 8*, 915-931. doi:10.1093/gbe/evw045

The Heliconius Genome Consortium. (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature, 487*, 94-98. doi:10.1038/nature11041

Conte, G. L., Arnegard, M. E., Peichel, C. L., & Schluter, D. (2012). The probability of genetic parallelism and convergence in natural populations. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 279*, 5039–5047.

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology, 23*, 3133-3157. doi:10.1111/mec.12796

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics, 27*, 2156-2158. doi:10.1093/bioinformatics/btr330

Delmore, K. E., Hubner, S., Kane, N. C., Schuster, R., Andrew, R. L., Camara, F., . . . Irwin, D. E. (2015). Genomic analysis of a migratory divide reveals candidate genes for migration and implicates selective sweeps in generating islands of differentiation. *Molecular Ecology, 24*, 1873-1888. doi:10.1111/mec.13150

Dincă, V., Lukhtanov, V. A., Talavera, G., & Vila, R. (2011). Unexpected layers of cryptic diversity in wood white *Leptidea* butterflies. *Nature Communications, 2*, e324.

Dincă, V., Wiklund, C., Lukhtanov, V. A., Kodandaramaiah, U., Noren, K., Dapporto, L., . . . Friberg, M. (2013). Reproductive isolation and patterns of genetic differentiation in a cryptic butterfly species complex. *Journal of Evolutionary Biology, 26*, 2095-2106. doi:10.1111/jeb.12211

Ellegren, H., Smeds, L., Burri, R., Olason, P., Backström, N., Kawakami, T., . . . Wolf, J. B. W. (2012). The genomics of species differentiation in *Ficedula* flycatchers. *Nature, 491*, 756-760.

Ellegren, H., & Wolf, J. B. W. (2017). Parallelism in genomic landscapes of differentiation, conserved genomic features and the role of linked selection. *Journal of Evolutionary Biology, 30*, 1516-1518. doi:10.1111/jeb.13113

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology, 14*, 2611–2620.

Feder, J. L., Flaxman, S. M., Egan, S. P., & Nosil, P. (2013). Hybridization and the build-up of genomic divergence during speciation. *Journal of Evolutionary Biology, 26*, 261-266. doi:10.1111/jeb.12009

Ferchaud, A. L., & Hansen, M. M. (2016). The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: three-spine sticklebacks in divergent environments. *Molecular Ecology, 25*, 238-259. doi:10.1111/mec.13399

Freese, A., & Fiedler, K. (2002). Experimental evidence for specific distinctness of the two butterfly taxa, *Leptidea sinapis* and *L. reali*. *Nota Lepidopterologica, 25*, 39-59.

Friberg, M. (2007). A difference in pupal morphology between the sibling species *Leptidea sinapis* and *L. reali* (Pieridae). *Nota lepidopterologica, 30*, 61-64.

Friberg, M., Leimar, O., & Wiklund, C. (2013). Heterospecific courtship, minority effects and niche separation between cryptic butterfly species. *Journal of Evolutionary Biology, 26*, 971-979. doi:10.1111/jeb.12106

Friberg, M., Olofsson, M., Berger, D., Karlsson, B., & Wiklund, C. (2008a). Habitat choice precedes host plant choice - niche separation in a species pair of a generalist and a specialist butterfly. *Oikos, 117*, 1337-1344.

Friberg, M., Vongvanich, N., Borg-Karlsson, A.-K., Kemp, D. J., Merilaita, S., & Wiklund, C. (2008b). Female mate choice determines reproductive isolation between sympatric butterflies. *Behavioral Ecology and Sociobiology, 62*, 873-886.

Friberg, M., & Wiklund, C. (2009). Host plant preference and performance of the sibling species of butterflies *Leptidea sinapis* and *Leptidea reali:* a test of the trade-off hypothesis for food specialisation. *Oecologia, 159*, 127-137. doi:10.1007/s00442-008-

Friberg, M., & Wiklund, C. (2010). Host-plant-induced larval decision-making in a habitat/host-plant generalist butterfly. *Ecology, 91*, 15-21.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv, 1207.3907*.

Gould, S. J. (1990). *Wonderful Life: The Burgess Shale and the Nature of History*. New York, USA: W. W. Norton & Co.

Harr, B. (2006). Genomic islands of differentiation between house mouse subspecies. *Genome Research, 16*, 730-737.

Hill, W. G., & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research, 8*, 269-294.

Irwin, D. E., Alcaide, M., Delmore, K. E., Irwin, J. H., & Owens, G. L. (2016). Recurrent selection explains parallel evolution of genomic regions of high relative but low absolute differentiation in a ring species. *Molecular Ecology, 25*, 4488-4507. doi:10.1111/mec.13792

Jiggins, C. D., & Martin, S. H. (2017). Glittering gold and the quest for Isla de Muerta. *Journal of Evolutionary Biology, 30*, 1509-1511. doi:10.1111/jeb.13110

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., . . . Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature, 484*, 55-61. doi:10.1038/nature10944

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution, 30*, 772-780. doi:10.1093/molbev/mst010

Kawakami, T., Wallberg, A., Olsson, A., Wintermantel, D., DeMiranda, J. R., Allsopp, M., . . . Webster, M. T. (2019). Substantial heritable variation in recombination rate on multiple scales in honeybees and bumblebees. *Genetics, [Early Online]*. doi:10.1534/genetics.119.302008

Keightley, P. D., Pinharanda, A., Ness, R. W., Simpson, F., Dasmahapatra, K. K., Mallet, J., . . . Jiggins, C. D. (2014). Estimation of the spontaneous mutation rate in *Heliconius melpomene*. *Molecular Biology and Evolution, 32*, 239-243.

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *Bioinformatics, 15*, 356.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution, 33*, 1870-1874.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transformation. *Bioinformatics, 26*, 589-595.

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature, 475*, 493-496.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data, P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*, 2078-2079. doi:10.1093/bioinformatics/btp352

Lohse, K. (2017). Come on feel the noise - from metaphors to null models. *Journal of Evolutionary Biology, 30*, 1506-1508. doi:10.1111/jeb.13109

Lorkovic, Z. (1993). *Leptidea reali* Reissinger 1989 (=*lorkovicii* Real, 1988), a new European species (Lepidoptera, Pieridae). *Natura Croatica, 2*, 1-26.

Lukhtanov, V. A., Dincă, V., Friberg, M., Sichova, J., Olofsson, M., Vila, R., . . . Wiklund, C. (2018). Versatility of multivalent orientation, inverted meiosis, and rescued fitness in holocentric chromosomal hybrids. *Proceedings of the National Academy of Sciences of the United States of America, 115*, E9610-E9619. doi:10.1073/pnas.1802610115

Lukhtanov, V. A., Dincă, V., Talavera, G., & Vila, R. (2011). Unprecedented within-species chromosome number cline in the wood white butterfly *Leptidea sinapis* and its significance for karyotype evolution and speciation. *BMC Evolutionary Biology, 11*, e109.

Lunter, G., & Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research, 21*, 936-939. doi:10.1101/gr.111120.110

Marec, F. (1996). Synaptonemal complexes in insects. *International Journal of Insect Morphology and Embryology, 25*, 205-233.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal, 17*, 10-12.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., . . . Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research, 23*, 1817-1828. doi:10.1101/gr.159426.113

Martin, S. H., Möst, M., Palmer, W. J., Salazar, C., McMillan, W. O., Jiggins, F. M., & Jiggins, C. D. (2016). Natural selection and genetic diversity in the butterfly *Heliconius melpomene*. *Genetics, 203*, 525-541. doi:10.1534/genetics.115.183285

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*, 1297-1303. doi:10.1101/gr.107524.110

Meyer, J. R., Dobias, D. T., Weitz, J. S., Barrick, J. E., Quick, R. T., & Lenski, R. E. (2012). Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science, 335*, 428-432.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., & Thomas, P. D. (2016). PANTHER version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research, 45*, D183-D189. doi:10.1093/nar/gkw1138

Nachman, M. W., & Payseur, B. A. (2012). Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*, 409-421. doi:10.1098/rstb.2011.0249

Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology, 18*, 375-402.

Pereira, R. J., Barreto, F. S., Pierce, N. T., Carneiro, M., & Burton, R. S. (2016). Transcriptome-wide patterns of divergence during allopatric evolution. *Molecular Ecology, 25*, 1478–1493.

Poelstra, J. W., Vijay, N., Bossu, C. M., Lantz, H., Ryll, B., Muller, I., . . . Wolf, J. B. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science, 344*, 1410-1414. doi:10.1126/science.1253226

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlovic, M., . . . Westram, A. M. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology, 30*, 1450-1477. doi:10.1111/jeb.13047

Renaut, S., & Dion-Cote, A.-M. (2016). History repeats itself: genomic divergence in copepods. *Molecular Ecology, 25*, 1417-1419.

Renaut, S., Grassa, C. J., Yeaman, S., Moyers, B. T., Lai, Z., Kane, N. C., . . . Rieseberg, L. H. (2013). Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications, 4*, 1827. doi:10.1038/ncomms2833

Renaut, S., Owens, G. L., & Rieseberg, L. H. (2014). Shared selective pressure and local genomic landscape lead to repeatable pat- terns of genomic divergence in sunflowers. *Molecular Ecology, 23*, 311-324.

Saitou, N., & Nei, M. (1987). The neighbour-joining method: a new method to reconstruct phylogenetic trees. *Molecular Biology and Evolution, 4*, 406-425.

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., . . . Widmer, A. (2014). Genomics and the origin of species. *Nature Reviews Genetics, 15*, 176-192. doi:10.1038/nrg3644

Šíchová, J., Ohno, M., Dincă, V., Watanabe, M., Sahara, K., & Marec, F. (2016). Fissions, fusions, and translocations shaped the karyotype and multiple sex chromosome constitution of the northeast-Asian wood white butterfly, *Leptidea amurensis*. *Biological Journal of the Linnean Society, 118*, 457-471.

Šíchová, J., Volenikova, A., Dincă, V., Nguyen, P., Vila, R., Sahara, K., & Marec, F. (2015). Dynamic karyotype evolution and unique sex determination systems in *Leptidea* wood white butterflies. *BMC Evolutionary Biology, 15*, 89. doi:10.1186/s12862-015-0375-4

Solovyev, V. I., Ilinsky, Y., & Kosterin, O. E. (2015). Genetic integrity of four species of *Leptidea* (Pieridae, Lepidoptera) as sampled in sympatry in West Siberia. *Comparative Cytogenetics, 9*, 299-324. doi:10.3897/CompCytogen.v9i3.4636

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics, 123*, 585-595.

Talla, V., Kalsoom, F., Shipilina, D., Marova, I., & Backström, N. (2017a). Heterogeneous patterns of genetic diversity and differentiation in European and Siberian chiffchaff (*Phylloscopus collybita abietinus / P . tristis*). *G3: Genes, Genomes, Genetics, 7,* 3983-3998.

Talla, V., Suh, A., Kalsoom, F., Dincă, V., Vila, R., Friberg, M., . . . Backström, N. (2017b). Rapid increase in genome size as a consequence of transposable element hyperactivity in wood-white (*Leptidea*) butterflies. *Genome Biology and Evolution, 9,* 2491-2505.

Tamura, K., Nei, M., & Kumar, S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences of the United States of America, 101,* 11030-11035.

Turner, T. L., Hahn, M. W., & Nuzhdin, S. V. (2005). Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology, 3,* 1572-1578.

Van Doren, B. M., Campagna, L., Helm, B., Illera, J. C., Lovette, I. J., & Liedvogel, M. (2017). Correlated patterns of genetic diversity and differentiation across an avian family. *Molecular Ecology, 26,* 3982-3997. doi:10.1111/mec.14083

Verovnik, R., & Glogovcan, P. (2007). Morphological and molecular evidence of a possible hybrid zone of *Leptidea sinapis* and *L. reali* (Lepidoptera: Pieridae). *European Journal of Entomology, 104,* 667-674. doi:10.14411/eje.2007.084

Vijay, N., Bossu, C. M., Poelstra, J. W., Weissensteiner, M. H., Suh, A., Kryukov, A. P., & Wolf, J. B. (2016). Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. *Nature Communications, 7,* 13195. doi:10.1038/ncomms13195

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution, 38,* 1358-1370.

Wiklund, C. (1977a). Behaviour in relation to female monogamy in *Leptidea sinapis* (*Lepidoptera*). *Oikos, 29,* 275-283.

Wiklund, C. (1977b). Oviposition, feeding and spatial separation of breeding andforaging habitats in a population of *Leptidea sinapis*. *Oikos, 28*, 56-68.

Wiklund, C., & Solbreck, C. (1982). Adaptive versus incidental explanations for the occurence of protandry in a butterfly, *Leptidea sinapis L. Evolution, 36*, 56-62.

Wolf, J. B., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics, 18*, 87-100. doi:10.1038/nrg.2016.133

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution, 19*, 395-420.

Zheng, X., Levine, D., Shen, J., Gogarten, S., Laurie, C., & Weir, B. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics, 28*, 3326-3328. doi:10.1093/bioinformatics/bts606

Zimmer, F., Harrison, P. W., Dessimoz, C., & Mank, J. E. (2016). Compensation of dosage-sensitive genes on the chicken Z chromosome. *Genome Biol Evol, 8*, 1233-1242.

Data accessibility

All raw sequence reads, the genome assembly and genotype files (.vcf and .bam) have been deposited in the European Nucleotide Archive (ENA) under accession number: PRJEB21838. In house developed scripts and pipelines are available at: https://github.com/venta380/leptidea-popgen (see also Supplementary Methods).

Author contributions

NB and VT designed research and lead the study. VT performed the analyses and wrote the manuscript together with NB. AJ and VT processed the raw sequencing data. VD, RV, MF and CW provided samples and advice on analysis and interpretation of results. All authors approved the final version of the manuscript before submission.

Tables

Table 1.
Summary of number of polymorphic sites (#SNPs), average pair-wise nucleotide diversity (θ_π) and Tajima's *D* estimates (T_D) for the three species (a) and all six populations individually (b) as estimated from non-overlapping 10 kb windows across the genome. The nucleotide diversity was significantly reduced on the Z-chromosome as compared to the autosome in all species / populations (Mann-Whitney U-tests, P-value range: $7.9*10^{-249}$ - $1.5*10^{-51}$)

a)

Species	**# SNPs**	**θ_π total**	**θ_π Autosomes**	**θ_π Z**	**T_D**
L. sinapis	8,514,582	0.0033±0.0015	0.0033±0.0015	0.0029±0.0012	-0.4±0.6
L. reali	4,209,882	0.0026±0.0014	0.0026±0.0014	0.0019±0.0012	0.3±0.8
L. juvernica	3,352,368	0.0016±0.0009	0.0017±0.0009	0.0015±0.0007	0.0±0.6

b)

Population	**# SNPs**	**θ_π total**	**θ_π Autosomes**	**θ_π Z**	**T_D**
L. sinapis (Swe)	6,087,199	0.0032±0.0016	0.0032±0.0016	0.0025±0.0014	-0.1±0.8
L. sinapis (Kaz)	5,308,021	0.0030±0.0016	0.0030±0.0016	0.0022±0.0013	0.1±0.8
L. sinapis (Spa)	5,810,369	0.0031±0.0015	0.0031±0.0015	0.0025±0.0012	-0.2±0.6
L. reali (Spa)	4,209,882	0.0026±0.0014	0.0026±0.0014	0.0019±0.0012	0.3±0.8
L. juvernica (Kaz)	3,818,134	0.0023±0.0012	0.0023±0.0012	0.0020±0.0009	0.1±0.6
L. juvernica (Ire)	2,262,560	0.0015±0.0009	0.0015±0.0009	0.0013±0.0007	0.8±0.8

Table 2.

Levels of genetic differentiation (F_{ST}) and absolute divergence (D_{XY}) across species and population comparisons. Summary statistics are given for all chromosomes combined (All) and the autosomes (Auto) and the Z-chromosome (Z), separately. Lsin = *Leptidea sinapis*, Lrea = *Leptidea reali*, Ljuv = *Leptidea juvernica*. Population abbreviations follow the standard in the main paper.

	F_{ST}			D_{XY} (*10^{-3})		
Pair	**All**	**Auto**	**Z**	**All**	**Auto**	**Z**
Lsin-Lrea	0.15±0.05	0.15±0.05	0.20±0.06	4.0±1.7	4.0±1.7	4.4±1.6
Lrea-Ljuv	0.28±0.08	0.28±0.08	0.34±0.08	3.8±1.5	3.8±1.5	4.3±1.5
Lsin-Ljuv	0.16±0.04	0.16±0.04	0.20±0.04	3.4±1.4	3.4±1.4	3.9±1.3
LsSwe-LsKaz	0.02±0.03	0.02±0.03	0.03±0.04	3.1±1.5	3.1±1.5	2.4±1.3
LsSwe-LsSpa	0.13±0.06	0.13±0.06	0.17±0.07	4.0±1.7	4.0±1.8	3.7±1.5
LsKaz-LsSpa	0.14±0.07	0.14±0.07	0.19±0.08	3.9±1.7	4.0±1.7	3.7±1.5
LjKaz-LjIre	0.09±0.06	0.09±0.06	0.11±0.05	1.8±0.9	1.8±0.9	1.7±0.8

Table 3.

Levels of absolute divergence (D_{XY} $*10^{-3}$) in relative differentiation (F_{ST}^{Z}) outliers as compared to all other genomic regions (Global) for the three species comparisons and the four intraspecific population comparisons. P-values represent Mann-Whitney U-tests for each comparison, respectively.

Pair	Outliers	Global	P-value
Lsin-Lrea	5.3±2.8	4.0±1.7	$8.8*10^{-29}$
Lrea-Ljuv	4.0±2.1	3.8±1.5	0.32
Lsin-Ljuv	3.6±1.8	3.4±1.4	0.082
LsSwe-LsKaz	3.4±1.8	3.1±1.5	$1.1*10^{-4}$
LsSwe-LsSpa	4.7±2.3	4.0±1.8	$2.1*10^{-12}$
LsKaz-LsSpa	4.7±2.2	4.0±1.7	$2.4*10^{-17}$
LjKaz-LjIre	1.7±1.2	1.8±0.9	$2.5*10^{-8}$

Figures

Figure 1.
Approximate distribution ranges and sampling locations of specimens. Pink distribution range = *L. sinapis*, violet range = *L. reali*, green range = *L. juvernica*. *L. sinapis* and *L. reali* are sympatric over the entire range of *L. reali*. *L. juvernica* and *L. sinapis* are sympatric in the major part of the distribution range but *L. juvernica* is mostly allopatric on Ireland and *L. sinapis* occurs in allopatry (with respect to *L. juvernica*) mainly in the southmost and northmost parts of the distribution range and in UK. Red circles = *L. sinapis* samples, blue circle = *L. reali* samples and green circles = *L. juvernica* samples. Note that the distribution ranges are inferred from previous sampling efforts and detailed knowledge about the exact ranges is not available, especially in the eastern parts. A *L. sinapis* male is shown in the top left corner. Note that *L. juvernica* and *L. reali* are virtually identical.

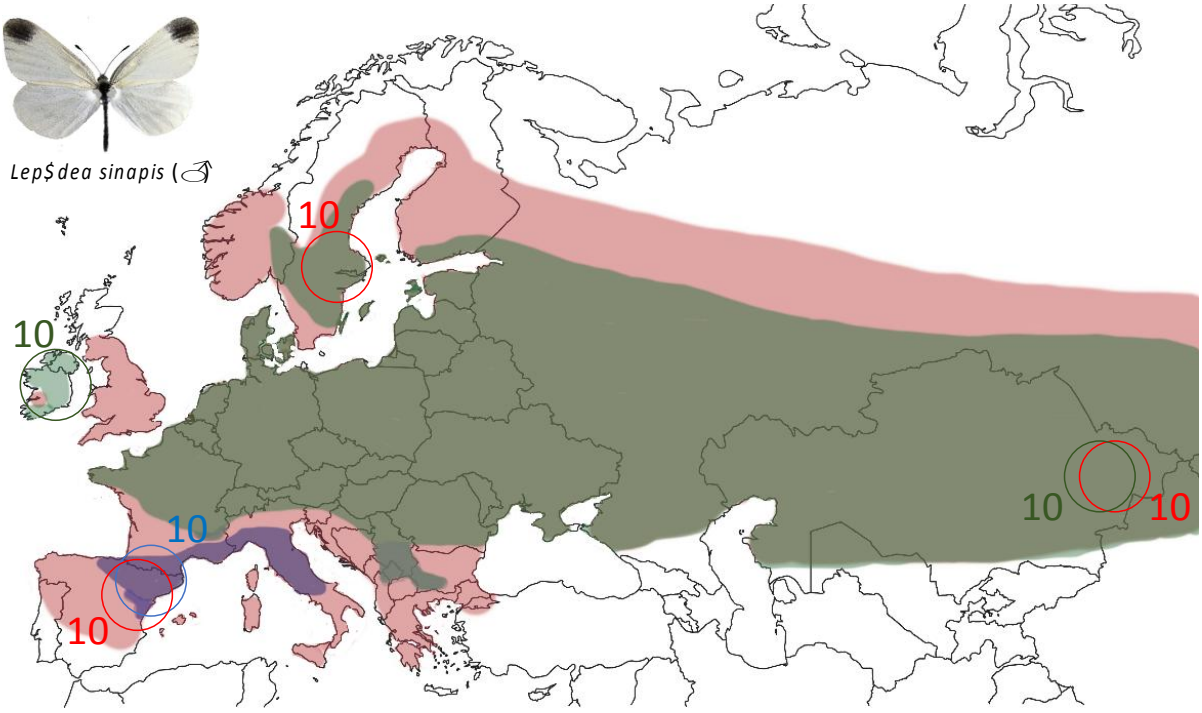


Figure 2.
Analyses of species and population relationships and changes in population size across time. A) A neighbor-joining phylogenetic tree based on the entire mitochondrial genome showing the relationship of all individuals in the *Leptidea* clade. *Bombyx mori*, *Heliconius melpomene* and *Phoebis sennae* were used as outgroups. Colour coding of individuals follows the standard throughout the paper. Bootstrap support values are given when 100%. The scale bar indicates the number of substitutions per site. B) Admixture proportions of individuals when all samples were analysed jointly for K = 3-5. Colour coding of individuals follows the standard throughout the paper. C) Error estimates for different K-values ranging from 1 to 11. The lowest error score was observed for K = 3, 4 and 5, corresponding to species separation (K = 3), additional structure within *L. juvernica* (K = 4) and additional structure within *L. sinapis* (K = 5). D) PCA plot showing the clustering of individuals based on the two principal components (PC1 and PC2) explaining most of the variance. E) Historical variation in effective population size as estimated by a PSMC-analysis.

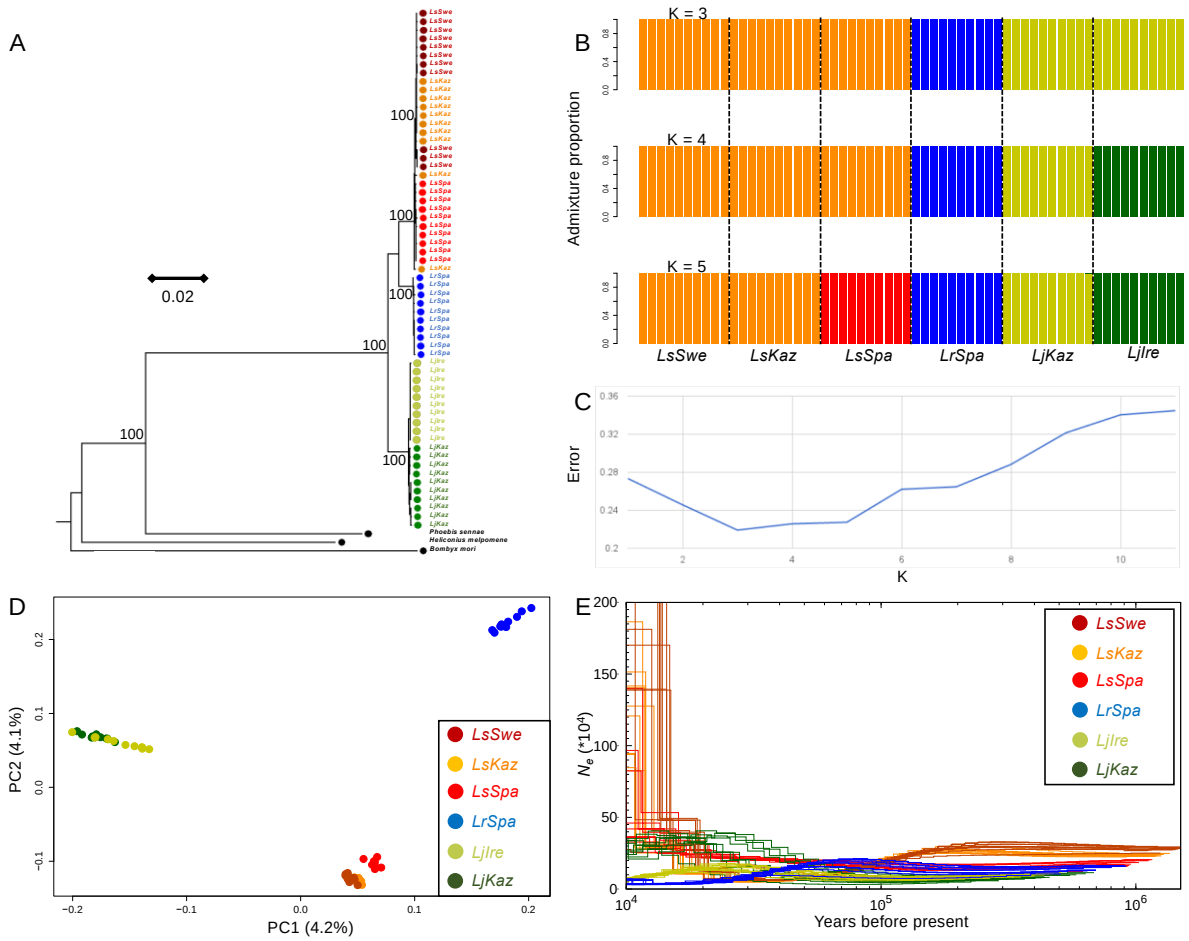


Figure 3.
Density distributions of genetic diversity (θ_π), calculated in non-overlapping 10-kb windows across the genome of the three *Leptidea* species (A) and for each of the six populations, separately (B).

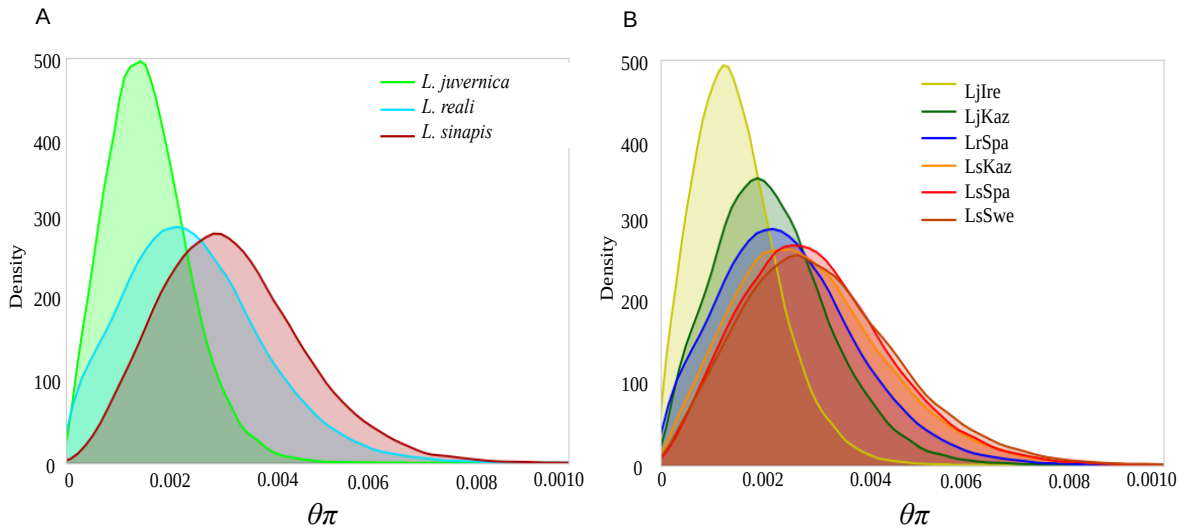


Figure 4.
Illustration of the regional variation in genetic differentiation (F_{ST}, top panel), absolute divergence (D_{XY}, second panel), genetic diversity (θ_{π}, third panel) and Tajima's *D* (*Taj D*, bottom panel) across the genome for the three species. The position along the genome (x-axis) is inferred from synteny with *H. melpomene*. Note that potential chromosomal rearrangements between *Leptidea* and *Heliconius* are not accounted for. Chromosomes are represented by grey and white blocks, the last block (far right) represents the Z-chromosome.

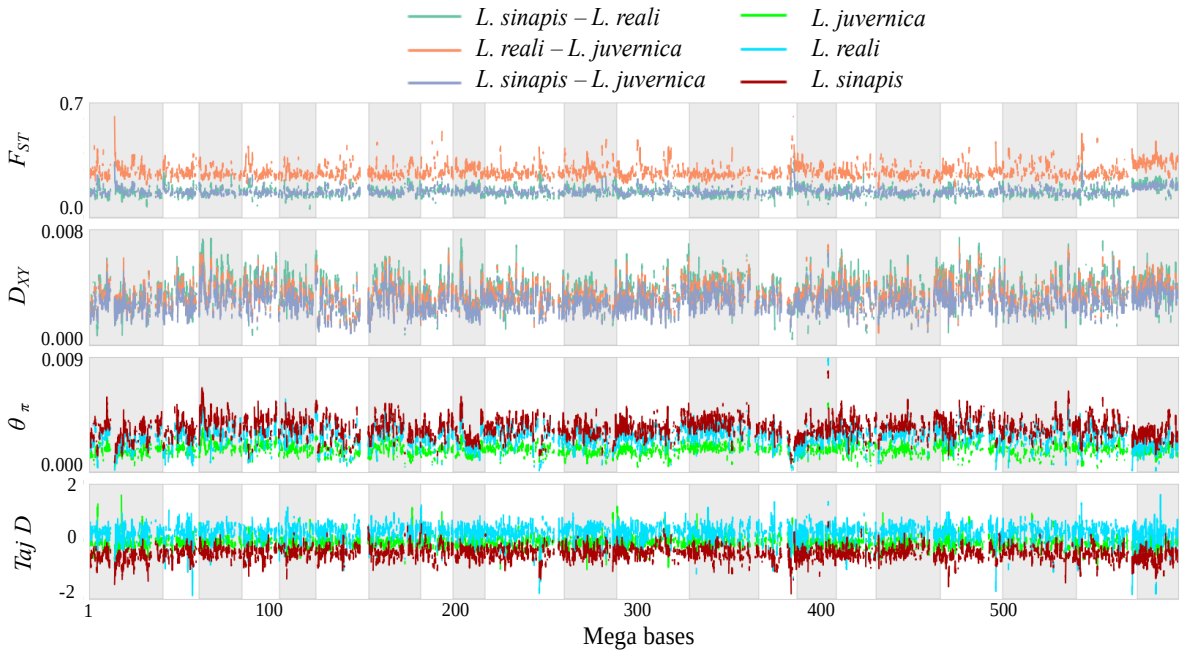


Figure 5.
Violin plots showing the distribution of absolute differentiation across 10 kb windows on autosomes (green) and the Z-chromosome (orange) for intraspecific population pairs (left part, bold face text) and species comparisons (right part, italics style text). Horizontal lines within each distribution indicate mean (dashed) and standard deviation (dotted).

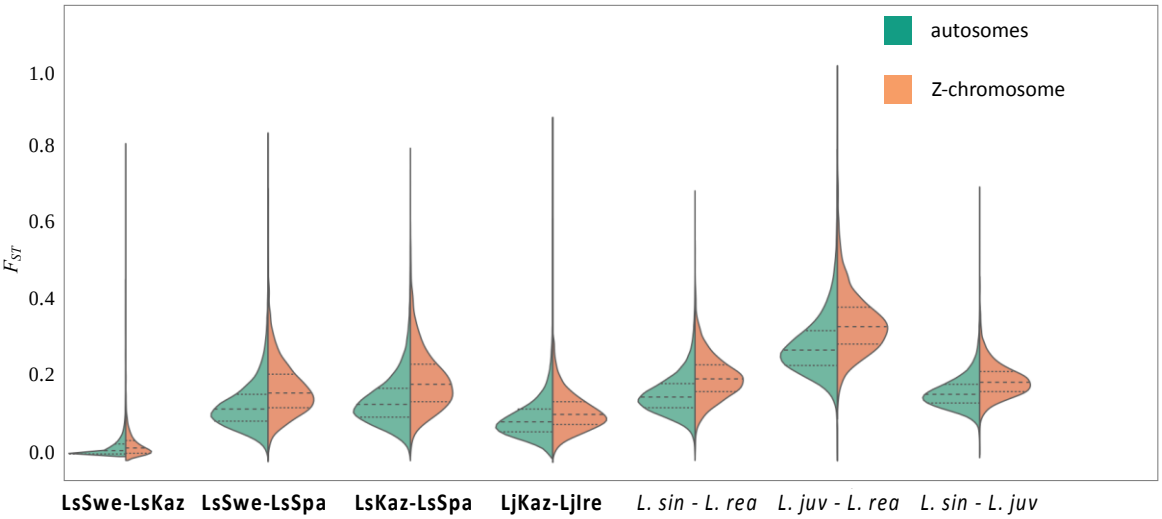
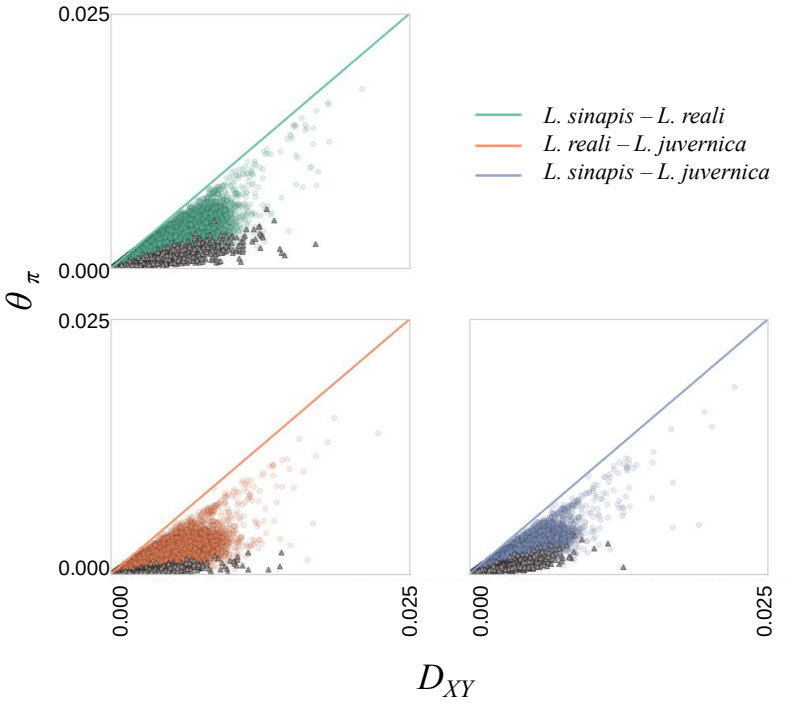


Figure 6.
Scatter plots illustrating the correlation between genetic diversity (θ_π, y-axis) and absolute divergence (D_{XY}, x-axis) across species (A) and population pairs (B). Significantly differentiated regions (top 1% of F_{ST}^{Z} windows) are indicated with grey triangles.

A



B