Whole-genomes from the extinct Xerces Blue butterfly reveal low diversity and long-term population decline

Toni de-Dios[1+], Claudia Fontsere[1+], Pere Renom[1+], Josefin Stiller[2], Laia Llovera[1], Marcela Uliano-Silva[3], Charlotte Wright[3], Esther Lizano[1,4], Arcadi Navarro[1,5], Robert K. Robbins[6], Mark Blaxter[3], Tomàs Marquès-Bonet[1,4,5,7]*, Roger Vila[1]*, Carles Lalueza-Fox[1]*

[1]Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), 08003 Barcelona, Spain

[2]Centre for Biodiversity Genomics, University of Copenhagen, DK-2100, Denmark

[3]Wellcome Sanger Institute, Hinxton, Saffron Walden CB10 1RQ, UK

[4]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain

[5] Catalan Institution of Research and Advanced Studies (ICREA), 08010 Barcelona, Spain

[6] Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20013-7012, USA

[7] CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), 08036 Barcelona, Spain

+ equally contributed

*Corresponding Authors

Carles Lalueza-Fox
Tomas Marquès-Bonet
Roger Vila

Email: carles.lalueza.fox@gmail.com

Abstract

The Xerces Blue (*Glaucopsyche xerces*) is considered to be the first butterfly to become extinct at global scale in historical times. It was notable for its chalky lavender wings with conspicuous white spots on the ventral wings. The last individuals were collected in their restricted habitat, in the dunes near the Presidio military base in San Francisco, in 1941. To explore the demographic history of this iconic butterfly and to better understand why it went extinct, we sequenced at medium coverage the genomes of four 80 to 100-year-old Xerces Blue specimens and seven historic specimens of its closest relative, the Silvery Blue (*G. lygdamus*). We compared these to a novel annotated genome of the Green-Underside Blue (*G. alexis*). Phylogenetic relationships inferred from complete mitochondrial genomes indicate that Xerces Blue was a distinct species that diverged from the Silvery Blue lineage at least 850,000 years ago. Using nuclear genomes, we show that both species experienced population growth during the MIS 7 interglacial period, but the Xerces Blue decreased to a very low effective population size subsequently, a trend opposite to that observed in the Silvery Blue. Runs of homozygosity in the Xerces Blue were significantly greater than in the Silvery Blue, suggesting a higher incidence of inbreeding. In addition, the Xerces Blue carried a higher proportion of derived, putatively deleterious amino acid-changing alleles than the Silvery Blue. These results demonstrate that the Xerces Blue experienced more than 100 thousand years of population decline, prior to its human-induced final extinction.

Keywords: Extinction, Ancient Genomics, Butterflies, Population Genomics, Xerces Blue

Introduction

The Xerces Blue butterfly (*Glaucopsyche xerces* (Boisduval))(1) was native to the coastal sand dunes of San Francisco in association with its preferred larval host plant, *Lotus scoparius* (Fabaceae) (2). It was notable for its iridescent blue colouration on the dorsal (upper) wing surface, and conspicuous, variable white spots on the ventral surface. With the growth of San Francisco and the destruction of sand dune habitats, the Xerces Blue became restricted to a few sites in what is now Golden Gate National Recreation Area. The last specimens were reportedly collected by entomologist W. Harry Lange on March 23, 1941, and the Xerces blue has never been seen flying again (2). It is considered the first butterfly to have been driven to extinction by human activities (2).

The Xerces Blue and the closely related Silvery Blue, *Glaucopsyche lygdamus* (Doubleday)*,* were recently proposed to be distinct species based on mtDNA data from a Xerces Blue museum specimen (3). However, two nuclear genes analysed (ribosomal 28S and histone H3) were invariable and genome-wide data were unavailable for the Xerces Blue, hampered by the inherent difficulties of retrieving genome-wide data from historical insect specimens (4, 5) and the absence of a suitable reference genome. The genus *Glaucopsyche* consists of 18 extant species distributed across the temperate regions of the northern hemisphere. To provide a relevant reference, we generated an annotated genome from the Palearctic Green-Underside Blue butterfly *Glaucopsyche alexis* (Poda). Using DNA extracted from five Xerces Blue and seven Silvery Blue (*Glaucopsyche lygdamus incognitus* Tilden) historical specimens from the vicinity of San

Francisco we generated whole genome resequencing data for both species and investigated their relationships and historical population genetics.

Results

Historic and modern butterfly genomes

Glaucopsyche alexis was chosen as a congeneric reference to compare the demographic histories of both the Xerces Blue and the Silvery Blue (Fig. 1). We generated a *G. alexis* reference genome from a male specimen collected in Alcalá de la Selva in Teruel (Spain). Its genome has a sequence length of 619,543,730 bp on 24 chromosomes – including the Z sex chromosome – and the mitochondrial genome (6). The genome sequence is biologically complete (BUSCO v5.1.2 Lepidoptera completeness 97.1%).

We extracted DNA from 12 historical specimens (5 *G. xerces*, 7 *G. lygdamus*) (Table S1). One Xerces Blue sample did not yield detectable DNA in two independent extractions. For each of the successful extracts we prepared a single library which was shotgun sequenced on the HiseqX Illumina platform. We mapped 124,101,622 and 184,084,237 unique DNA reads of Xerces Blue and Silvery Blue, respectively, against the *G. alexis* reference genome (Table S2). The DNA reads exhibited typical ancient DNA features, such as short mean read length (ranging from 47.55 to 67.41 bases on average, depending on the specimen (Fig. S1)) and post-mortem deamination patterns at the 5' and 3' ends (Table S2) (Fig. S2). The historical genomes covered 49.3% (Xerces Blue) and 55.2% (Silvery Blue) of the *G. alexis* reference genome, largely because repetitive chromosomal regions cannot be confidently assessed with short, ancient DNA sequence reads. To estimate the mappable fraction of the reference *G. alexis* genome, we randomly fragmented it to 50 to 70 nucleotides and mapped the generated fragments back to the complete genome. An average of 57.8% of the *G. alexis* genome was covered with these read lengths (Fig. S3). We suggest that reduced coverage from the historical specimens may be due to genomic divergence of *G. xerces* and *G. lygdamus* from the *G. alexis* reference (Fig. S4).

The sex of the specimens was determined by differential coverage of the Z chromosome (females are the heterogametic sex in the Lepidoptera and show reduced coverage on the Z chromosome). As listed in the original museum records, we found one Silvery Blue and two Xerces Blue females (Table S2). Inter-individual comparisons suggested no close kinship link among the studied individuals.

Phylogenetic relationships

Maximum likelihood phylogenetic inference using whole mitochondrial genomes (15,268 nucleotides) showed that the Xerces Blue specimens form a monophyletic clade, as do the Silvery Blue specimens (Fig. 1A). We inferred a time-calibrated Bayesian phylogenetic tree from protein-coding genes (11,028 nucleotides) of the mitochondrial DNA genomes of Xerces Blue, Silvery Blue, and 12 related butterflies in Polyommatinae. The Xerces Blue and the Silvery Blue are recovered as sister taxa with high support (posterior probability=1). Because there are no known fossils to calibrate the time since divergence, we first used a molecular clock that spanned the range of rates frequently used for arthropod mitochondrial genes (1.5-2.3% divergence/Ma). This yielded an origin of this subgroup of Polyommatinae at 10.26 Ma (7.16-13.76 Ma 95% HPD interval) and divergence of the Xerces Blue from the Silvery Blue at 849,000 years ago (560,895-1,159,512 years 95% HPD interval, Fig. 1B). A second estimate based on larger-scale fossil-based calibrations (7) placed the origin of the subgroup at ca. 33 Ma (8) with the subsequent divergence of the Xerces Blue and Silvery Blue at 2.60 Ma (2.14-3.08 Ma 95% HPD interval, Figure 1B). The recent speciation of Xerces and Silvery blue is not obviously due to infection with the *Wolbachia*, as no evidence of infection of the sampled specimens with this alpha-proteobacterium is detected in the raw read data.

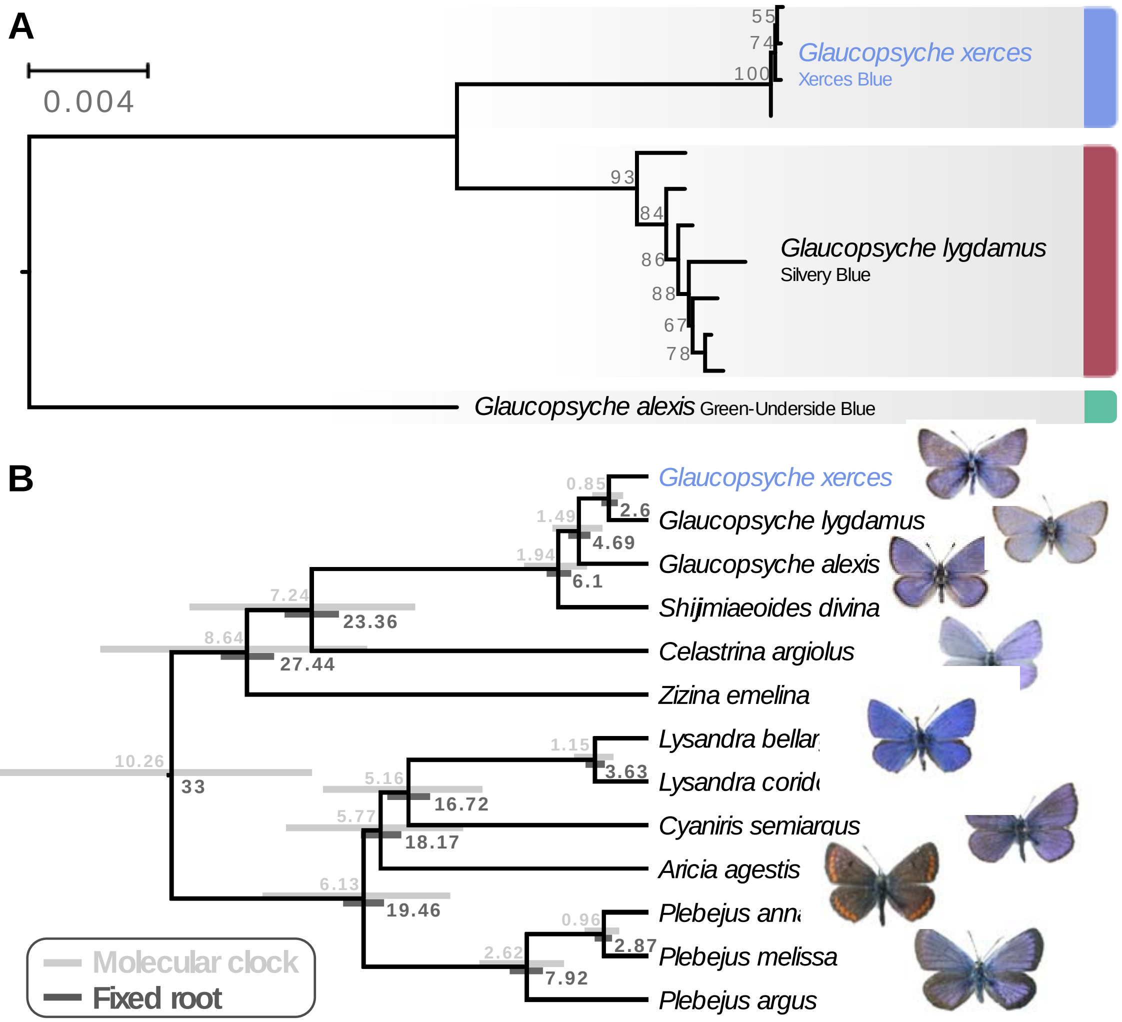

Fig. 1. Phylogenetic placement of the Xerces Blue. A: Maximum likelihood tree from whole mitochondrial genomes of Xerces Blue, Silvery Blue and Green-Underside Blue. Node labels are bootstrap support values. B: Time-calibrated phylogeny from Bayesian inference using mitochondrial protein-coding genes of Xerces Blue and related butterflies. Node values show median age estimates from dating analysis with a molecular clock (above nodes) or from fixing the age of the root (below nodes). Bars are 95% HPD intervals for node ages.

Principal Component Analysis (PCA) using PCAngsd (9) and nuclear genome polymorphisms for the three *Glaucopsyche* species supports the relationships among them; the historical specimens are equally distant to *G. alexis* in the first PC, explaining 56.16% of the variance (Fig. S5). The second PC separates the Xerces Blue from the Silvery Blue specimens.

Demographic history and diversity

We used the Pairwise Sequentially Markovian coalescent (PSMC) algorithm (10) to evaluate the demographic histories of both butterfly species, first exploring the two specimens with highest coverage (L05 and L13) (Fig. 2). We found an increase in effective population size in both species that is roughly coincident with the interglacial Marine Isotopic Stage (MIS) 7 (approximately from 240,000 to 190,000 years ago (11)). After this timepoint the trends differ. We estimated a continuous decrease in Xerces Blue population size in parallel to the Wisconsin Glacial Episode, which started about 75,000 years ago. However, the Silvery Blue does not appear to have been similarly affected by this event, suggesting different adaptive strategies to cope with cooling temperatures and/or food plant availability.

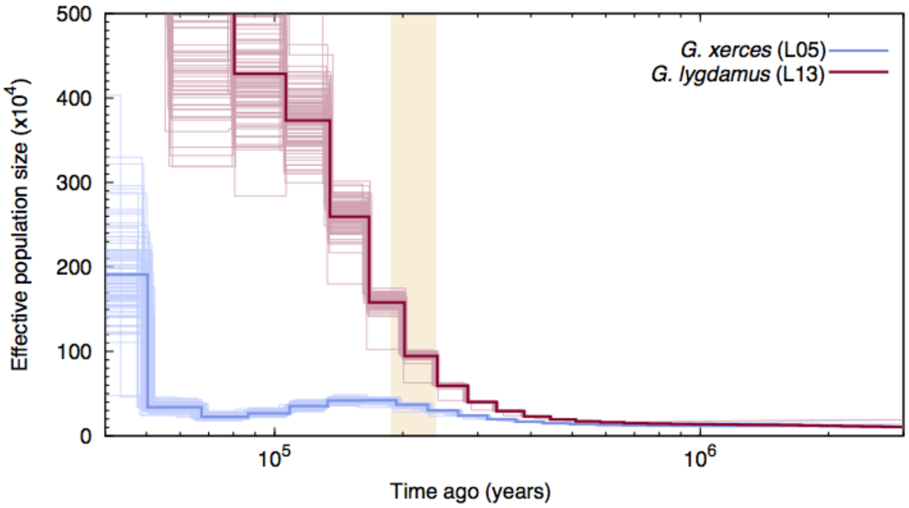

Fig. 2. PSMC plot of one Xerces Blue (L05) and one Silvery Blue (L13) specimen. The two samples are those with higher average coverage. Individual PSMC plots were bootstrapped 100 times each (lighter lines). One year of generation time and a mutation rate of μ=1.9x10^{-9} were used. The peak of the MIS 7 interglacial is marked in yellow.

We generated PSMC curves from the remaining lower-coverage individuals and down-sampled data from specimen L05 to 50% and 75% of the total coverage to explore the effects of coverage on estimation of heterozygous sites. Although there was a reduction in the effective population size estimates, as expected, the temporal trajectories in lower-coverage individuals were similar to their respective, higher-coverage Xerces Blue and Silvery Blue references (Fig. S6).

We subsequently explored the heterozygosity of each individual and found that Xerces Blue had 22% less heterozygosity on average than the Silvery Blue, a difference that is statistically significant (T-test; p=0.0072) (Fig. S7, Table S1). We searched for runs of homozygosity (RoH) that can indicate the existence of inbreeding in a dwindling population. The total fraction of the genome presenting RoH, although limited, is much higher in Xerces Blue (up to 6% of the genome) than in Silvery Blue, especially in short RoH of size between 100 and 500 kb (Fig. 3 and Fig. S7), consistent with background inbreeding. The limited presence of long RoH discards consanguinity as a common scenario in Xerces Blue.

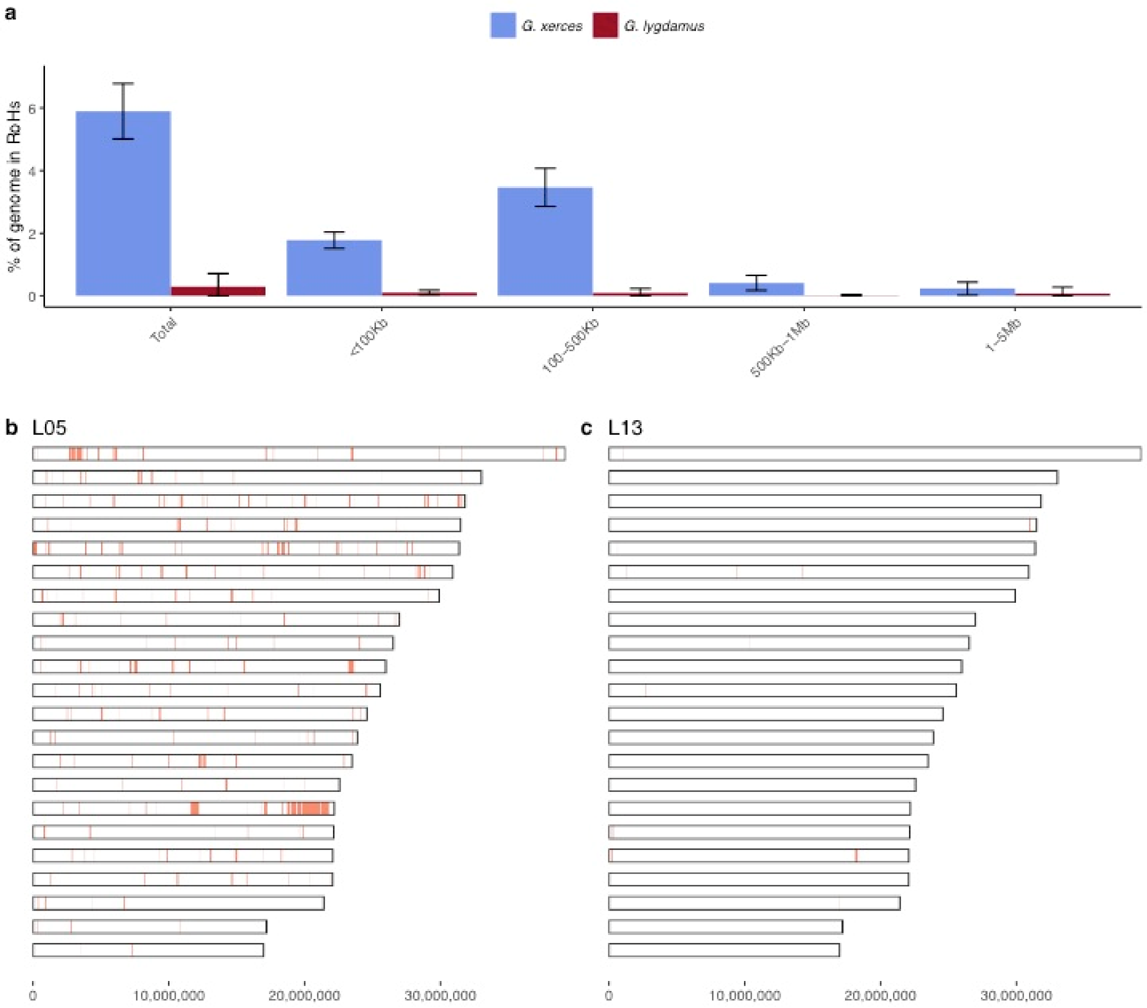

Fig. 3. Runs of Homozygosity (RoH) in the genomes of Xerces Blue and Silvery Blue. a: Percentage of the autosomal genome in RoH by size bins: very short RoH (<100 Kb), short RoH (100-500 Kb), intermediate RoH (500Kb-1Mb) and long (1-5Mb). Short RoH reflect LD patterns, intermediate size RoH describe background inbreeding due to genetic drift and long RoH appear in the case of very recent inbreeding due to consanguinity. Error bars show the standard deviation. b: Distribution of RoH in the autosomal genome of a Xerces specimen, L05 c) Distribution of RoH in the autosomal genome of a Silvery specimen L13.

We identified amino acid-changing alleles that may be suggestive of a deleterious genetic load associated with long-term low population numbers in the Xerces Blue. The average Ka/Ks ratio is higher in Xerces Blue than in Silvery Blue; the former also carries a higher fraction of nonsense and functionally high-to-moderate effect variants in homozygosis and RoHs with an increased concentration of high-to-moderate effect variants (Fig. 4), as predicted with a functional prediction toolbox, SnpEff (12). We could not assess the presence of non-synonymous mutations in genes associated to changes in wing pigmentation patterns.

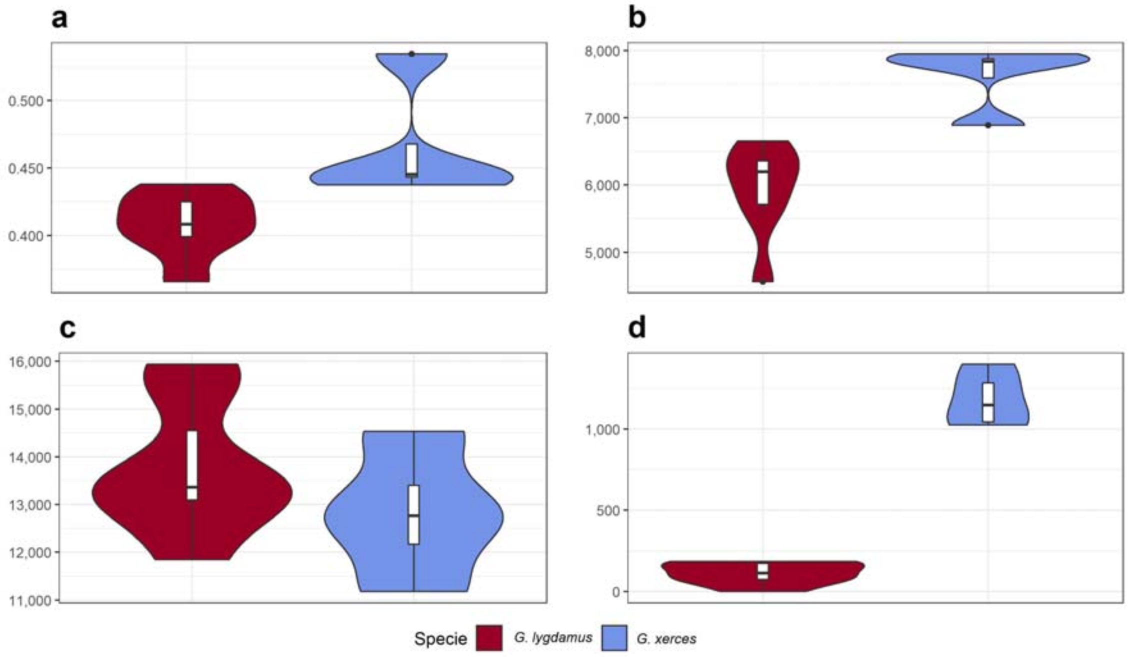

Fig. 4. Functional effect prediction on the fixed amino acid-changing alleles observed in Xerces Blue and Silvery Blue. a: Wide genome Ka/Ks ratio comparison. b: High-to-moderate effect variant comparison in homozygous sites. c: High-to-moderate effect variant comparison in heterozygous sites. d: Presence of high-to-moderate variants in regions of the genome in RoH. Error bars show the standard deviation.

Wolbachia analysis

Wolbachia are endosymbiotic alpha-proteobacteria that are present in about 70% of butterfly species and induce diverse reproductive alterations, including genetic barriers when two different strains infect the same population or when two populations – one infected and one uninfected – meet. As potential evidence for a reproductive barrier promoting the separation of Xerces Blue and Silvery Blue, we searched for *Wolbachia* DNA reads in our specimens, taking advantage of the high coverage and the shotgun approach, but failed to detect a solid presence (between 0 and 0.02% of DNA reads not attributable to our butterflies, depending on the specimen) of this bacterium.

Pigmentation genes

We detected two fixed variants in Xerces Blue not found in the Silvery Blue: the first is a T to A change in the gene *Wnt4*, corresponding to an intronic / downstream variant (FR990046.1: 19485537) and the second is an intronic A to T change in the *Wnt11* gene (FR990050.1: 16205523). Although Wnt4 and Wnt11 genes are not known to be involved in butterfly wing colour pattern (13), it can hypothesized that those two variants can be involved in changes at regulatory elements affecting phenotypical differences between both species.

Discussion

We have used a modern reference genome and ancient DNA genome sequence data from museum specimens to explore the relationships and historical population genetic history of an extinct butterfly, the Xerces Blue. Based upon a near-complete mtDNA genome from a Xerces Blue specimen, Grewe et al. (2021) (3) proposed that the Xerces Blue and the Silvery Blue were distinct species. We confirm this finding using full mitochondrial genomes and extensive nuclear genomic data from multiple specimens. Given the lack of evidence for *Wolbachia* infection, a detailed analysis of genomic architectures could help identify barriers to introgression between these species.

We provide evidence for low population size in Xerces Blue, correlated with low genetic variation, a higher proportion of runs of homozygosity and increased frequency of deleterious, amino acid-changing alleles (14–16). However, there was no genetic evidence of recent inbreeding, which sometimes occurs in critically endangered species (17,18). Our analyses indicate that the Xerces Blue had experienced a severe demographic decline for tens of thousands of years, likely associated with changing climatic factors. Thus, the destruction of the Xerces Blue habitat by humans was likely the final blow in the extinction process. We suggest that endangered insects with demographic traits indicative of long-term low effective population size should be considered to be especially vulnerable to extinction.

The generation of these paleogenomes is the first step towards understanding the specific adaptations of Xerces Blue butterfly. Maybe the most conspicuous of these was a pattern of white spots, usually without black ocelli, on their ventral wing surfaces. Genes such as *Wnt*, *optix* and *cortex* are associated with pigmentation wing patterns in butterflies (19–23). It may be informative to identify amino acid differences in these and other genes and to undertake functional studies, such as those based on CRIPSR-Cas 9 knock-outs, in extant relatives to elucidate relevant phenotypic and genomic characteristics of this extinct butterfly. However, in our preliminary screening in these pigmentation genes, no amino acid changes as compared to the Silvery Blue genomes have been discovered, raising the question if potential differences could be in fact in regulatory regions. However, it is likely that this question would need to further increase the coverage of the paleogenomes generated. Irrespective of this, our study further demonstrates the value of ancient DNA in museum specimens for evolutionary studies at a population scale.

Glaucopsyche alexis is a good reference genome to study the transition from mutualism with ants to parasitism on ants within subtribe Glaucopsychina; the latter group includes the remarkable *Phengaris* butterflies, which are ant-parasites in the last instars and are currently endangered in Europe (mainly due to their highly specialised strategy) (24). The *G. alexis* reference genome (coupled with those of the Xerces Blue and Silvery Blue) will help to understand the evolution from an exclusively herbivorous diet in later instars to a aphytophagous one and the biochemical synthesis of molecules used in the ant-parasitic relationship.

Materials and Methods

Historical butterfly specimens

The Xerces Blue specimens belong to the Barnes collection deposited at the Smithsonian National Museum of Natural History. Two of them were collected on April 26th, 1923. The Silvery Blue specimens were mostly collected between 1927 and 1948, in Haywood City, Santa Cruz, Oakland, San José, Fairfax and Marin County (these locations surround San Francisco Bay).

DNA extraction and sequencing of historical specimens of Xerces Blue and Silvery Blue

All DNA extraction and initial library preparation steps (prior to amplification) were performed in the dedicated clean lab, physically-isolated from the laboratory used for post-PCR analyses. Strict protocols were followed to minimize the amount of human DNA in the ancient DNA laboratory, including the wearing a full body suit, sleeves, shoe covers, clean shoes, facemask, hair net and double gloving. All lab surfaces consumables, disposables, tools and instruments are wiped with bleach and ethanol, and UV irradiated before and after use.

DNA extraction was performed from 12 abdominal samples of Xerces Blue and Silvery Blue. One ml of digestion buffer (final concentrations: 3 mM CaCl2, % SDS, 40 mM DTT, 0.25 mg/ml proteinase K, 100 mM Tris buffer pH 8.0 and 100 mM NaCl) was added to each crushed butterfly residue, including an extraction blank, and incubated at 37 °C overnight (24h) on rotation (750-900 rpm). Next, DNA extraction was continued following the method proposed by Dabney et al. (2013)(25). Remaining butterfly sample was then pelleted by centrifugation in a bench-top centrifuge for 2 min at maximum speed (16,100 × g). The supernatant was added to 10 mL of binding buffer (final concentrations: 5 M guanidine hydrochloride, 40% (vol/vol) isopropanol, 0.05% Tween-20, and 90 mM sodium acetate (pH 5.2)) and purified on a High Pure Extender column (Roche). DNA extracts were eluted with 45 μL of low EDTA TE buffer (pH 8.0) and quantified using a Qubit instrument.

Following extraction, 100-200 ng of DNA extract was converted into Illumina sequencing libraries following the BEST protocol(26). Each library was amplified by PCR using two uniquely barcoded primers. After index PCR, libraries were purified with a 1.5x AMPure clean (Beckman Coulter) and eluted in 25 μl of low EDTA TE buffer (pH 8.0). Libraries were quantified using BioAnalyzer and sequenced by HiSeq 4000 (Illumina).

Glaucopsyche alexis genome sequencing

The *G. alexis* genome was sequenced at the Sanger Institute as part of the Darwin Tree of Life Project [https://wellcomeopenresearch.org/articles/6-27] following the extraction, sequencing and assembly protocols developed for Lepidoptera. Data are available in INSDC under BioProject PRJEB43798 and genome assembly accessions GCA_905404095.1 (primary haplotype) and GCA_905404225.1 (secondary, alternate haplotype).

Glaucopysche alexis annotation

Prior to annotation, a de novo repeat library was made using RepeatModeller v2.0.1 (27) and was then provided to RepeatMasker v4.1.1 to detect and mask repetitive sequences (28). Protein-coding genes and transcripts were predicted using BRAKER2 v2.0.6 (29), using protein hints which were generated using the ProtHint protein mapping pipeline v2.6 (30) with the reference protein sequences for arthropods from OrthoDB (31) as homology evidence.

Xerces Blue and Silvery Blue mapping and variant calling

The ancient DNA reads were clipped using AdapterRemoval v2.2 (32), sequencing adapters were removed. Only reads longer than 25bp were kept. Filtered reads were mapped against the *G. alexis* assembly with Burrows-Wheeler Aligner (BWA) (33), using the *backtrack* algorithm, setting no trimming, disabling seed (-l 1024), increasing stringency for edit distance (-n 0.01), and allowing opening of 2 gaps (33, 34) Duplicated reads were removed using -tools MarkDuplicates (35). Mapped reads with mapping quality below 25 were removed using Samtools (36).

Basic mapping statistics were generated using Qualimap 2 (37). The resulting reads were examined with PMDtools and MapDamage2 to assess the degradation rate of the data, which is a sign of authenticity (38,39). We detected the presence of typical aDNA-damaged bases at the end of reads and *pmd score* distribution that could be attributed to an authentic museum sample (38) (Fig. S8). To avoid problems in the next steps, we trimmed 2 nt from each read end using BamUtil trimbam (40). Bedtools was to assess genome coverage across the reference, using windows of 1mbp for the nuclear fraction of the genome (41). To account for the reference genome mappability, we split the reference genome in reads of 60bp and re-mapped them against the reference. Coverage and mappability are displayed using Circos (42).

We used snpAD v0.3.2 (43), a program developed for genotype calling in ancient specimens since it infers priors for each sample separately to account for DNA damage patterns. The mapped sequences were transformed from bam-format into snpAD-format files, priors for base composition estimated, and genotypes were called using standard settings. The VCFs were combined and concatenated with CombineVariants and GatherVcfs from GATK v3.5 (44) and filtered with vcftools v0.1.12b (45) to keep only sites within the mappable fraction of the genome with minimum read depth of 2, max read depth of 30, genotype quality > 30, maximum missingness of 0.6, minor allele frequency of 5% and excluding indels and multiallelic sites.

Genotype likelihoods were obtained with ANGSD (46) version 0.916 using the GATK model with the following parameters for all the samples: -uniqueOnly 1 -remove_bads 1 -

only_proper_pairs 1 -trim 10 -C 50 -baq 1 -minInd 5 -skipTriallelic 1 -GL 2 -minMapQ 30 -doGlf 2 -doMajorMinor 1 -doMaf 2 -minMaf 0.05 -SNP_pval 1e-6.

Sex determination

We calculated the depth of coverage of each sample in windows of 10kb using Mosdepth (47). Next, we normalized each window coverage with the mean coverage of each individual to exclude the influence of sequence effort and plotted the density of normalized mean coverage in the autosomes and the sex chromosome (chrZ) to genetically identify the sex of each sample (half of the average coverage is expected in the chrZ for females).

Mitochondrial phylogenetic tree and divergence dating

Complete mitochondrial genomes were annotated with MitoFinder (48) and aligned with MAFFT l-ins-i. We first investigated phylogenetic relationships among *Glaucopsyche* individuals by analyzing the entire mitochondrial genome. We used IQ-TREE2 (49) to select the best fitting substitution model for each partition and merge similar partitions and build a maximum likelihood tree and assessed support with 1000 ultrafast bootstrap replicates.

To infer a time-calibrated phylogeny, we selected one individual each of Xerces Blue and Silvery Blue in addition to 12 related Polyommatinae. We extracted the sequences for protein-coding genes, aligned each with the codon-aware aligner MACSE, and concatenated all. We used BEAST2 (50) with the bModelTest package to perform phylogenetic site model averaging for each of the merged partitions. Because there is no accepted molecular clock rate for butterflies and no fossils to apply in this part of the phylogeny, we used two strategies to apply time constraints to the analysis. First, we used two published molecular clock rates for the mitochondrial COX1 gene (1.5% divergence/Ma estimated for various invertebrates (51), and the 'standard' insect mitochondrial clock 2.3% divergence/Ma (51). We applied a strict clock with a normal prior set up to span the 1.5-2.3% range with the 95% HPD interval (mean=1.9%, sigma=0.00119). Second, we borrowed the age of the most recent common ancestor of our sampled taxa from fossil-calibrated analyses across butterflies, which has been estimated to ~33 Ma(7, 52). We fixed the root age to 33 Ma and allowed the remaining node ages to be estimated using a strict clock. Analyses were run twice from different starting seeds for 10 million MCMC generations and trees were sampled every 1000 generations. Runs were checked for convergence with Tracer and all effective sample size (ESS) values were >200. Runs were combined with LogCombiner, after removing the first 30% of topologies as burn-in, and a maximum credibility tree was generated with TreeAnnotator (50). Phylogenetic analyses were performed on the National Life Science Supercomputing Center - Computerome 2.0 (www.computerome.dk).

Xerces Blue and Silvery Blue population histories

We used the Pairwise Sequentially Markovian Coalescent (PSMC) model to explore the demographic history of both species (Xerces Blue and Silvery Blue). We obtained a consensus fastq sequence of the mappable fraction of the genome for each autosomal chromosome (total of 22 chromosomes of *G. alexis* assembly). Only positions with a depth of coverage above 4X and below 15X were kept. Posteriorly a PSMC was built using the following parameters: -N25 -t15 -r5 -p "28*2+3+5". We used 1 year for the generation time and a mutation rate of $1.9x10^{-9}$, estimated in *Heliconius melpomene* (53). Considering that calling consensus sequences from low coverage samples (< 10x) can underestimate heterozygous sites (54), and given the different coverage between samples, we corrected by False Negative Rate the samples with coverage lower than the coverage of L005 (for Xerces Blue) and L013 (for Silvery Blue), as recommended by the developers of the software, so that all samples are comparable with each other. However, since in our dataset we do not reach a coverage > 20x, we acknowledge that we are not capturing the whole diversity and thus our PSMC might infer lower historical effective population sizes.

Population stratification and average genome heterozygosity

Principal Component Analysis (PCA) was performed using PCAngsd after obtaining genotype likelihoods with ANGSD including all individuals and plotted with R. To assess global levels of heterozygosity, the unfolded SFS was calculated for each sample separately using ANGSD and realSFS with the following quality filter parameters: - uniqueOnly 1 - remove_bads 1 -only_proper_pairs 1 -trim 10 -C 50 -baq 1 -minMapQ 30 -minQ 30 -setMaxDepth 200 - doCounts 1 -GL 2 -doSaf 1.

Runs of Homozygosity (RoH)

RoH were called based on the density of heterozygous sites in the genome using the implemented Hidden Markov Model (HMM) in bcftools roh (55) with the following parameters: -G30 --skip-indels --AF-dflt 0.4 --rec-rate $1e^{-9}$ from the mappable fraction of the genome with the filtered VCF file. We kept the RoH with a phred score > 75. We divided the RoH into different size bins: very short RoH (<100 kb), short RoH (100-500 kb), intermediate RoH (500 kb-1Mb) and long (1-5 Mb or > 5Mb). Short RoH reflect LD patterns, intermediate size RoH describe background inbreeding due to genetic drift and long RoH appear in the case of recent inbreeding (56). Results were plotted in R v 3.6.3 (57).

Deleterious load and fixed mutations

We used the *G. alexis* annotations to create a SNPeff database that we used to annotate our callings generated with angsd and predict their putative effect (12). Additionally, we took a more in-depth view to genes which could alter the coloration patterns in the individuals (19–23). First we located those genes in our annotation with BLAST and their homologs in other butterfly species (58), setting an E-value lower than 0.001 and an Identity value above 60%. Then, the coordinates were called using GATK v3.7 UnifiedGenotyper. Variants were filtered for a indels and minimum Genotype Quality of 30 using vcftools (45). Variants were kept regardless of their coverage. A variant was considered as fixed in a specie if it is covered in at least 2 individuals of each specie, do not present heterozygous genotypes, and when one of the species present all their genotypes calls as homozygous for the derived allele while in the other are homozygous of the ancestral allele. Additionally, we checked the coverage in the after mention genes using bedtools (41) (Fig. S9). The results were displayed using ggplot2 (59).

Screening of intracellular parasites

To explore the presence of relevant intracellular parasites that could explain their speciation process (60), we collapsed unique reads from the butterfly-free sequences and removed from the dataset low complexity sequences using Prinseq (61). Afterwards, we applied kraken2 to assign reads against standard contaminant databases (bacteria, archaea, fungi, protozoa and viral)(62).

Pigmentation genes

To find possible amino acid-changing variants that could explain phenotypical differences between *G. lygdamus* and *G. xerces*, we have taken a more definite look at genes associated to colour patterns in butterflies; namely *optix*, *cortex* and *Wnt* genes (19-23). We have first identified these genes in our annotation using a homology-based approach with BLAST algorithm tblastn (57). We used the homologs in the butterfly species *Papilo machaon, Heliconius erato, Zerene cesonia, Danaus plexippus* and *Bombyx mori* as proxies (63-66). We set a E-value lower than 0.001 and an Identity value above 60% for the tblastn results (Table S.3). Variants were then called at the gene coordinates using GATK v3.7 UnifiedGenotyper (67). We filtered the variants for indels and minimum Genotype Quality of 30 using vcftools v.1.14, but we kept them regardless of their coverage (45). A variant was considered as fixed in a species if it was covered in at least 2 individuals, does not present heterozygous genotypes, and if either Xerces Blue or Silvery Blue show all their genotypes calls as homozygous for the derived allele while being homozygous for the ancestral allele in the other species. Additionally, we checked the coverage in the target genes using bedtools v2.2.2 (41) (Fig. S9).

Acknowledgments

C.L.-F. is supported by a PGC2018-0955931-B-100 grant (MCIU/AEI/FEDER, UE) of Spain; T.M.-B. is supported by funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 864203), BFU2017-86471-P (MINECO/FEDER, UE), “Unidad de Excelencia María de Maeztu”, funded by the AEI (CEX2018-000792-M), Howard Hughes International Early Career, Secretaria d’Universitats i Recerca and CERCA Programme del Departament d’Economia i Coneixement de la Generalitat de Catalunya (GRC 2017-SGR-880); R.V. is supported by grant PID2019-107078GB-I00, funded by MCIN/AEI/ 10.13039/501100011033, and by GRC 2017-SGR-991 (Generalitat de Catalunya). We are grateful to the SCIENCE Faculty at University of Copenhagen for free access to Computerome 2.0. The genetic data generated is publicly available; the accession numbers for the Xerces Blue and Silvery Blue genomes reported in this study are in the European Nucleotide Archive (ENA): PRJEB47122. This research was funded in whole, or in part, by the Wellcome Trust Grants 206194 and 218328 (MU, MB). For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Competing Interest Statement: The authors declare no competing interest.

Author Contributions: C.L.-F., R.V., R.R. and P.R. conceived the project. R.R. studied and sampled the specimens. L.L. and E.L. performed experimental work. T.d-D., C.F., J.S., M.U.-S., C.W. and P.R. undertook different computational analyses. C.L.-F., T.M.-B., A.N. and M.B. coordinated different computational teams. J.S. worked in visualization. C.L.-F. and R.R. wrote the manuscript with input from all coauthors.

References

1. J. A. Boisduval, Lépidoptères de la Californie. *Ann. Soc. Ent. France* **21**, 275–324 (1852).

2. J. C. Downey, W.H. Lange, Analysis of Variation in a recently extinct polymorphic lycaenid butterfly, Glaucopsyche Xerces. *Bull. South Calif. Acad. Sci.* **55**, 153–170 (1956).

3. F. Grewe, M. R. Kronforst, N. E. Pierce, C. S. Moreau, Museum genomics reveals the Xerces blue butterfly (*Glaucopsyche xerces*) was a distinct species driven to extinction . *Biol. Lett.* **17**, 20210123 (2021).

4. P. F. Thomsen, *et al.*, Non-destructive sampling of ancient insect DNA. *PLoS One* **4**, e5048 (2009).

5. M. Staats, *et al.*, Genomic Treasure Troves: Complete Genome Sequencing of Herbarium and Insect Museum Specimens. *PLoS One* **8**, e69189–e69189 (2013).

6. J.C. Hinojosa Galisteo, R. Vila, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium, The genome sequence of the green-underside blue, *Glaucopsyche alexis* (Poda, 1761). *Wellcome Open Res.* **6**, 274 (2021).

7. M. Espeland, *et al.*, A Comprehensive and Dated Phylogenomic Analysis of Butterflies. *Curr. Biol.* **28**, 770-778.e5 (2018).

8. N. Chazot, *et al.*, Priors and Posteriors in Bayesian Timing of Divergence Analyses: The Age of Butterflies Revisited. *Syst. Biol.* **68**, 797–813 (2019).

9. J. Meisner, A. Albrechtsen, Inferring population structure and admixture proportions in low-depth NGS data. *Genetics* **210**, 719–731 (2018).

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

11. C. L. Batchelor, *et al.*, The configuration of Northern Hemisphere ice sheets through the Quaternary. *Nat. Comm.* **10**, 3713 (2019).

12. M. Coon, *et al.*, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).

13. J. Fenner, *et al.*, Wnt Genes in Wing Pattern Development of Coliadinae

Butterflies. *Front. Ecol. Evol.* **8**, 197 (2020).

14. Z. A. Szpiech, *et al.*, Long Runs of Homozygosity Are Enriched for Deleterious Variation. *Am. J. Hum. Genet.* **93**, 90–102 (2013).

15. D. Spielman, B. W. Brook, R. Frankham, Most species are not driven to extinction before genetic factors impact them. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 15261–15264 (2004).

16. E. Palkopoulou, *et al.*, Complete Genomes Reveal Signatures of Demographic and Genetic Declines in the Woolly Mammoth. *Curr. Biol.* **25**, 1395–1400 (2015).

17. T. van der Valk, D. Díez-Del-Molino, T. Marques-Bonet, K. Guschanski, L. Dalén, Historical Genomes Reveal the Genomic Consequences of Recent Population Decline in Eastern Gorillas. *Curr. Biol.* **29**, 165-170.e6 (2019).

18. D. Díez-del-Molino, F. Sánchez-Barreiro, I. Barnes, M. T. P. Gilbert, L. Dalén, Quantifying Temporal Genomic Erosion in Endangered Species. *Trends Ecol. Evol.* **33**, 176–185 (2018).

19. L. Zhang, R. D. Reed, Genome editing in butterflies reveals that spalt promotes and Distal-less represses eyespot colour patterns. *Nat. Comm.* **7**, 11769 (2016).

20. L. Zhang, A. Mazo-Vargas, R. D. Reed, Single master regulatory gene coordinates the evolution and development of butterfly color and iridescence. *Proc. Natl. Acad. Sci. USA* **114**, 10707–10712 (2017).

21. A. Mazo-Vargas, *et al.*, Macroevolutionary shifts of WntA function potentiate butterfly wing-pattern diversity. *Proc. Natl. Acad. Sci.* **114**, 10701–10706 (2017).

22. J. Fenner, *et al.*, Wnt Genes in Wing Pattern Development of Coliadinae Butterflies. *Front. Ecol. Evol.* **8**, 197 (2020).

23. T. Das Banerjee, S. K. Shan, A. Monteiro, optix is involved in eyespot development via a possible positional information mechanism. *bioRxiv* (2021) https:/doi.org/10.1101/2021.05.22.445259.

24. L. V. Ugelvig, R. Vila, N. E. Pierce, D. R. Nash, A phylogenetic revision of the Glaucopsyche section (Lepidoptera: Lycaenidae), with special focus on the Phengaris-Maculinea clade. *Mol. Phyl. Evol.* **61**, 237–243 (2011).

25. J. Dabney, *et al.*, Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. USA* **110**, 15758–15763 (2013).

26. C. Carøe, *et al.*, Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* **9**, 410–419 (2018).

27. J.M. Flynn, et al., RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451-9457 (2020).

28. A.F.A. Smit, R. Hubley, P. Green. RepeatMasker. Published at: http://www.repeatmasker.org (2013).

29. T. Brůna, K. J. Hoff, A. Lomsadze, M. Stanke, M. Borodovsky, BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinforma.* **3,** lqaa108 (2021).

30. T. Brůna, A. Lomsadze, M. Borodovsky, GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics Bioinforma.* **2,** lqaa026 (2020).

31. E. V Kriventseva, *et al.*, OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).

32. M. Schubert, S. Lindgreen, L. Orlando, AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).

33. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

34. M. Schubert, *et al.*, Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* **13**, 178 (2012).

35. Broad Institute, Picard. Web page: http://broadinstitute.github.io/picard/

36. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

37. K. Okonechnikov, A. Conesa, F. García-Alcalde. Qualimap2: advanced multi-scale quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292-294.

38. P. Skoglund, *et al.*, Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc. Natl. Acad. Sci. USA.* **111**, 2229–2234 (2014).

39. H. Jónsson, A. Ginolhac, M. Schubert, P. L. F. Johnson, L. Orlando, MapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013).

40. G. Jun, M.K. Wing, G.R. Abecasis, H.M. Kang, An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* **25**, 918–925 (2015).

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

42. M. Krzywinski, *et al.*, Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

43. K. Prüfer, SNPAD: An ancient DNA genotype caller. *Bioinformatics* **34**, 4165–

4171 (2018).

44. G. A. Van der Auwera, *et al.*, From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1-33 (2013).

45. P. Danecek, *et al.*, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

46. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014).

47. B. S. Pedersen, A. R. Quinlan, Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).

48. R. Allio, *et al.*, MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol. Ecol. Resour.* **20**, 892–905 (2020). 2

49. L. T. Nguyen, H. A. Schmidt, A. Von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

50. A. J. Drummond, A. Rambaut, BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).

51. S. P. Quek, S. J. Davies, T. Itino, N. E. Pierce, Codiversification in an ant-plant mutualism: Stem texture and the evolution of host use in Crematogaster (Formicidae: Myrmicinae) inhabitants of Macaranga (Euphorbiaceae). *Evolution* **58**, 554–570 (2004).

52. M. Wiemers, N. Chazot, C. W. Wheat, O. Schweiger, N. Wahlberg, A complete time-calibrated multi-gene phylogeny of the european butterflies. *Zookeys* **2020**, 97–124 (2020).

53. S. H. Martin, *et al.*, Natural selection and genetic diversity in the butterfly *Heliconius melpomene*. *Genetics* **203**, 525–541 (2016).

54. C. Sarabia, B. vonHoldt, J. C. Larrasoaña, V. Uríos, J. A. Leonard, Pleistocene climate fluctuations drove demographic history of African golden wolves (*Canis lupaster*). *Mol. Ecol.*, mec.15784 (2021).

55. V. Narasimhan, *et al.*, BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**, 1749–1751 (2016).

56. F. C. Ceballos, P. K. Joshi, D. W. Clark, M. Ramsay, J. F. Wilson, Runs of homozygosity: windows into population history and trait architecture. *Nat. Rev. Genet.* **19**, 220–234 (2018).

57. R. Ihaka, R. Gentleman, R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.* **5**, 299–314 (1996).

58. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

59. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016).

60. A. Telschow, P. Hammerstein, J. H. Werren, The effect of Wolbachia versus genetic incompatibilities on reinforcement and speciation. *Evolution* **59**, 1607–1619 (2005).

61. R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).

62. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

63. M. A. Supple, *et al.*, Genomic architecture of adaptive color pattern divergence and convergence in Heliconius butterflies. *Genome Res.* **23**, 1248–1257 (2013).

64. X. Li, *et al.*, Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies. *Nat. Commun.* **6**, 8212 (2015).

65. L. Gu, *et al.*, Dichotomy of Dosage Compensation along the Neo Z Chromosome of the Monarch Butterfly. *Curr. Biol.* **29**, 4071-4077.e3 (2019).

66. J. Ou, *et al.*, Transcriptomic analysis of developmental features of Bombyx mori wing disc during metamorphosis. *BMC Genomics* **15**, 820 (2014).

67. A. McKenna, *et al.*, The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

Supplementary Tables

Table S1: List of historical specimens analyzed in this study.

Table S2: General metrics of the historical genomes, mappability, heterozygosity estimates and number of RoH per specimen.

Table S3: Table S3: Chromosomal identification coordinates in the *G. alexis* annotation of the main pigmentation genes previously described in Lepidoptera.