Whole-genomes from the extinct Xerces Blue butterfly can help identify declining insect species

Toni de-Dios [1,2+], Claudia Fontsere [1,3+], Pere Renom [1+], Josefin Stiller [3], Laia Llovera[1], Marcela Uliano-Silva[4], Alejandro Sánchez-Gracia[5], Charlotte Wright[4], Esther Lizano [1,6], Berta Caballero[7], Arcadi Navarro [1,8], Sergi Civit [9], Robert K. Robbins[10], Mark Blaxter[4], Tomàs Marquès-Bonet [1,6,8,11*], Roger Vila [1*], Carles Lalueza-Fox [1,7*]

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

2Institute of Genomics, University of Tartu, Riia 23B, 51010 Tartu, Estonia

3Section for Evolutionary Genomics, The Globe Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen K, Denmark

3Centre for Biodiversity Genomics, University of Copenhagen, DK-2100, Denmark

4Wellcome Sanger Institute, Hinxton, Saffron Walden CB10 1RQ, UK

5Departament of Genetics, Microbiology and Statistics-Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona 08028, Spain

6Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain

7Museu de Ciències Naturals de Barcelona, 08019, Barcelona, Spain

8 Catalan Institution of Research and Advanced Studies (ICREA), 08010 Barcelona, Spain

9Departament of Genetics, Microbiology and Statistics-Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona 08028, Spain

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

11 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

Emails: (T.d.D) tonidedios94@gmail.com; (C.F.) claudia.fontsere@sund.ku.dk; (P.R.) cogitoergomitto@gmail.com; (J.S.) josefinstiller@gmail.com; (A.S.G.) elsanchez@ub.edu; (C.W.) cw22@sanger.ac.uk; (E.L.) esther.lizano@upf.edu; (B.C.) bcaballerolo@bcn.cat; (A.N). arcadi.navarro@upf.edu (S.C.) svives@ub.edu; (R.K.R.) ROBBINSR@si.edu; (M.B.) mb35@sanger.ac.uk; (T.M.-B) tomas.marques@upf.edu; (R.V.) roger.vila@csic.es; (C.L.-F) carles.lalueza.fox@gmail.com;

Running Head: Genomes of the extinct Xerces Blue butterfly

Keywords: Extinction; Museomics; Butterflies; Population Genomics; Xerces Blue; Genomic Erosion

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. We sequenced the genomes of four 80 to 100-year-old Xerces Blue, and seven historical and one modern 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, both species experienced population growth during the Eemian 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 and deleterious load in the Xerces Blue were significantly greater than in the Silvery Blue, suggesting a higher incidence of inbreeding. These signals of population decline observed in Xerces Blue could be used to identify and monitor other insects threatened by human activities, whose extinction patterns are still not well known.

Introduction

The Xerces Blue butterfly (*Glaucopsyche xerces)*[1] was native to the coastal sand dunes of San Francisco in association with the common Deerwood (*Acmispon glaber),* which was the preferred food source for larval stage[2]. It was notable for its iridescent blue colouration on the dorsal (upper) wing surface, and conspicuous, variable white spots on the ventral surface[3]. 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[3]. It is considered the first butterfly to have been driven to global extinction by human activities[3].

The Xerces Blue and the closely related Silvery Blue (*Glaucopsyche lygdamus*) were recently proposed to be distinct species based on mtDNA data from a single Xerces Blue specimen[4]. 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[5,6] 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*[7]. Using DNA extracted from five Xerces Blue and seven Silvery Blue (*Glaucopsyche lygdamus*) historical specimens from the vicinity of San Francisco, and also from a modern Silvery Blue male from Canada, we generated whole genome resequencing data for both species and investigated their relationships and historical population genetics.

Results

Historic and modern butterfly genomes

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 1 and 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 (Fig. S2 and S3). 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.

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 (Fig. S4 – S7). 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 (Table S2). 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. S8). The annotation of genes located in those unrecoverable regions provided a putative list of 14 nuclear genes with diverse functions obtained from BLAST, that should be further explored to understand the uniqueness of the extinct species (Table S4).

Phylogenetic relationships

Maximum likelihood phylogenetic inference using whole mitochondrial genomes 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 Analysis and 12 related butterflies in Polyommatinae subfamily, revealing high support for the sister group relationship (posterior probability=1). We found the specie *Shijimiaeoides (Sinia) divina* inside the *Glaucopsyche* clade, in agreement with previous phylogenetic studies[8]. 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). Our dated analysis yielded an origin of this subgroup of Polyommatinae at 12.4 Ma (8.82-16.27 Ma 95% HPD interval) and divergence of the Xerces Blue from the Silvery Blue at 900,000 years ago (0.61-1.19 Ma 95% HPD interval, Fig. 1B). A second estimate based on larger-scale fossil-based calibrations[9] fixed the origin of the subgroup to ca. 33 Ma[10], inferred the subsequent divergence of the Xerces Blue and Silvery Blue to 2.40 Ma (1.95-2.73 Ma 95% HPD interval, Fig 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.

Principal Component Analysis (PCA) using PCAngsd[11] 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 52.81% of the variance (Fig. S9). 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[12] 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 7 (approximately from 240,000 to 190,000 years ago[13]). 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, both the modern and the historical Silvery Blue do not

appear to have been negatively affected by this event (Fig. S10), suggesting different adaptive strategies to cope with cooling temperatures and/or food plant availability.

Second, 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. S10).

We subsequently explored the heterozygosity of each individual and found that Xerces Blue had 22% less heterozygosity on average than the Silvery Blue historical samples, a difference that is statistically significant (T-test; p=0.0072) (Fig. S8, Table S2). 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. S11), consistent with background inbreeding. The limited presence of long RoH discards consanguinity as a common scenario in Xerces Blue.

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 homozygosity and RoHs with an increased concentration of high-to-moderate effect variants (Fig. 4), as predicted with a functional prediction toolbox, SnpEff[14].

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; to our knowledge, this is the first ancient genome ever generated from an extinct insect. Based upon a near-complete mtDNA genome from a Xerces Blue specimen, Grewe et al. (2021)[4] 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, the recent speciation of Xerces Blue and Silvery Blue seems unrelated to cytoplasmic incompatibility cause by this endosymbiont[15,16]; a detailed analysis of genomic architectures could help identify barriers to introgression between these species.

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 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 [17–19]. However, there was no genetic evidence of recent inbreeding.

Inbreeding genetic signals in the form of long chromosomal sections with no variation sometimes occur in critically endangered species [20,21] and in extinct species such as the last Mammoths from Wrangel Island[22] or the Altai Neanderthal[23]. The PSMC shows a continuous low effective population size for Xerces Blue; demographic declines are also seen in some extinct species, including Wrangel Mammoths[19] but not in others such as the Woolly Rhino that showed a pre-extinction demographic stability and relatively low inbreeding signals[24]. In many endangered species there is little concordance between genome diversity, population sizes and conservation status[21];

this decoupling was also observed in the genomes of the extinct passenger pigeon that despite being one of the world's most numerous vertebrates, showed a surprisingly low genetic diversity[25]. Despite being notoriously abundant, insects, and in particular butterflies, are very sensitive to climate fluctuations; therefore, we suggest that insects with observations of demographic traits indicative of long-term low effective population size such as those found in Xerces Blue should be considered to be especially vulnerable to extinction events.

Our study further demonstrates the value of museum insect specimens for estimating temporal changes in genetic diversity at a population scale[26]. Despite being notoriously abundant, insects, and in particular butterflies, are very sensitive to climate fluctuations. We suggest that insects with genetic observations of long-term low effective population size such as those found in Xerces Blue should be considered to be especially vulnerable to extinction events. However, being the insect numbers usually very high, it is likely that their genomic signals of extinction could be different to those described in vertebrates in many cases. Therefore, this is a subject that should be further explored with genomic data from other declining insects.

Methods

Historical butterfly specimens

The Xerces Blue specimens analysed 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) (Table S1).

DNA extraction and sequencing of Xerces Blue and Silvery Blue specimens

All DNA extraction and initial library preparation steps (prior to amplification) were performed in a 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, as well as frequent bleach cleaning of benches and instruments. DNA extraction was performed from 12 abdominal samples of historical Xerces Blue and Silvery Blue, as well as a modern Silvery Blue specimen from Canada. Experimental procedures are described in detail in the Supplementary Material.

Glaucopsyche alexis genome sequencing and annotation

Glaucopsyche alexis was chosen as a congeneric reference to compare the demographic histories of both the Xerces Blue and the Silvery Blue. 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. The genome sequence is biologically complete (BUSCO Lepidoptera completeness 97.1%)[27]. The *G. alexis* genome was sequenced at the Sanger Institute as part of the Darwin Tree of Life Project following the extraction, sequencing and assembly protocols developed for Lepidoptera[7].

Xerces Blue and Silvery Blue mapping and variant calling

The ancient DNA reads were clipped using AdapterRemoval2[28], and only reads longer than 25bp were kept. Filtered reads were mapped against the *G. alexis* assembly with Burrows-Wheeler Aligner (BWA)[29], with parameters optimised for the analysis of aDNA (Supplementary Materials). Basic mapping statistics were generated using Qualimap2[30] (Table S2). We used bedtools[31] to assess genome coverage across the reference using windows of 1mbp for the nuclear fraction of the genome (Fig. S4 – S6),

as well as depth of coverage (Fig. S7), read length (Fig. S1) and edit distance distribution (Fig. S8). Authenticity of the sequences was assessed by characterising aDNA damage patterns with pmdtools[32] and MapDamage2[33] (Fig S2 and S3).

We used snpAD[34], a program for genotype calling in ancient specimens. 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[35] and filtered with vcftools[36] to keep only sites within the mappable fraction of the genome previously obtained 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[37] 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

The sex of the specimens was determined by differential coverage of the Lepidopteran Z chromosome (females are the heterogametic sex in the Lepidoptera and show reduced coverage on the Z chromosome) (Table S2).

Mitochondrial phylogenetic tree and divergence dating

Haploid variants were called using bcftools[38] with a ploidy of 1, filtering low quality indels and variants, after which a consensus sequence was exported. We downloaded 14 complete mitochondrial genomes for Polyommatinae from NCBI (Table S3).

All mitochondrial genomes were annotated with MitoFinder[39] using *Shijimiaeoides divina* as the reference. The 11 protein-coding genes were aligned with the codon-aware aligner MACSE[40] and the ribosomal rRNAs were aligned with MAFFT l-ins-i[41]. We first investigated phylogenetic relationships among five *G. xerces* and eight *G. lygdamus* individuals, with *G. alexis* as the outgroup. We used IQ-TREE2[42] to select the best fitting nucleotide substitution model for each partition and merge similar partitions[43], built a maximum likelihood tree and assessed support with 1000 ultrafast bootstrap replicates[44].

To infer a time-calibrated phylogenetic hypothesis, we selected one individual of Xerces Blue (L003) and Silvery Blue (RVcoll10-B005) and analysed with 13 other Polyommatinae species. We used BEAST2[45] with the bModelTest[46] 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[47], and the 'standard' insect mitochondrial clock 2.3% divergence/Ma[48]). We applied a strict clock with a normal prior set up to span 1.5-2.3% 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[10,49]. 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 the BEAST2 package LogCombiner[50], after removing the first 10% 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[12] to explore the demographic history of both butterfly species. 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 1.9×10^{-9}, estimated in *Heliconius melpomene*[51]. Considering that calling consensus sequences from low coverage samples (<10x) can underestimate heterozygous sites[52], 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[11] after obtaining genotype likelihoods with ANGSD including all individuals. To assess global levels of heterozygosity, the unfolded SFS was calculated for each sample separately using ANGSD[37] 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[38] roh 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 > 85. 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[53].

Deleterious load

We used the *G. alexis* annotations to create a SNPeff database that we used to annotate our callings. Using SNPeff[14] again and the set of variants discovered by angsd, we predicted the putative effect of those variant in the analysed individuals (Table S2). In addition to wide genome mutations, we specifically focused on mutations present in homozygosis, heterozygosis and the previously annotated RoH.

Unrecoverable regions

To further explore how the genomic divergence can influence our genome reconstruction success, we undertook a similar approach as the genome of the Christmas Island rat[54], and explored the chromosomal regions in the *G. alexis* reference that were significantly depleted of Xerces DNA reads. We used bedtools[31] and some in-home bash scripting to calculate the mean coverage per gene of the *G. alexis* genome for Xerces Blue sequencing DNA reads. We first used bedtools' algorithms *bamtobed* and *genomecov* to estimate the genome-wide per-site coverage of the reference genome in these two species. Then, we extracted the coordinates of all protein coding genes from the annotation file (gff file) and used the intersect to estimate the average coverage of each protein coding gene. We performed a

functional analysis of all genes uncovered in *G. alexis*, excluding those that are present in *G. lygdamus* with more than 5x coverage (as we were looking for evolutionary novelties in the Xerces Blue lineage alone) using profile- IntersProScan[55] and sequence similarity-based (blasp) searches[56].

Wolbachia screening

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[15]. 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. First, we collapsed unique reads from the butterfly-free sequences with BBmap[57] and removed from the dataset low complexity sequences using Prinseq[58]. Afterwards, we used Kraken2[59] to assign reads against the standard plus human Kraken2 database (bacteria, archaea, fungi, protozoa and viral). The historical specimens did not display enough reads assigned to Wolbachia for us to suspect of the presence of the bacteria in those samples (Table S6).

Acknowledgments

C.L.-F. is supported by a PID2021-124590NB-100 grant (MCIU/AEI/FEDER, UE) of Spain; T.M.-B. is supported by funding from the European Research Council (ERC) (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, and 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. 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., A.S.G., M.U.-S., C.W., S.C. 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., J.S. and R.R. wrote the manuscript with input from all coauthors.

Data Accessibility

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. Data on *G. alexis* are available in INSDC under BioProject PRJEB43798 and genome assembly accessions GCA_905404095.1 (primary haplotype) and GCA_905404225.1 (secondary, alternate haplotype).

References

1. Boisduval, J. A. Lépidoptères de la Californie. *Ann. Soc. Ent. Fr.* **21**, 275–324 (1852).
2. Tilden, J. W. San Francisco's vanishing butterflies. *Lepid. News* **10**, 133–1445 (1956).
3. Downey, J. C. Analysis of Variation in a recently extinct polymorphic lycaenid butterfly, Glaucopsyche Xerces. *Bull. So. Calif. Acad. Sci.* **55**, 153–170 (1956).
4. Grewe, F., Kronforst, M. R., Pierce, N. E. & Moreau, C. S. Museum genomics

reveals the Xerces blue butterfly (Glaucopsyche xerces) was a distinct species driven to extinction . *Biol. Lett.* **17**, 20210123 (2021).

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

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

7. Hinojosa Galisteo, J. C. & Vila, R. The genome sequence of the green-underside blue, Glaucopsyche alexis (Poda, 1761). *Wellcome Open Res.* **6**, 274 (2021).

8. Lukhtanov, V. A. & Gagarina, A. V. Molecular Phylogeny and Taxonomy of the Butterfly Subtribe Scolitantidina with Special Focus on the Genera Pseudophilotes, Glaucopsyche and Iolana (Lepidoptera, Lycaenidae). *Insects* **13** (12), 1110 (2022).

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

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

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

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

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

14. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).

15. Telschow, A., Hammerstein, P. & Werren, J. H. The effect of Wolbachia versus

genetic incompatibilities on reinforcement and speciation. *Evolution* **59**, 1607–1619 (2005).

16. Sucháčková Bartoňová, A. *et al.* Wolbachia affects mitochondrial population structure in two systems of closely related Palaearctic blue butterflies. *Sci. Rep.* **11**, 1–14 (2021).

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

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

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

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

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

22. Rogers, R. L. & Slatkin, M. Excess of genomic defects in a woolly mammoth on Wrangel island. *PLoS Genet.* **13**, e1006601 (2017).

23. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).

24. Lord, E. *et al.* Pre-extinction Demographic Stability and Genomic Signatures of Adaptation in the Woolly Rhinoceros. *Curr. Biol.* **30**, 3871-3879.e7 (2020).

25. Murray, G. G. R. *et al.* Natural selection shaped the rise and fall of passenger pigeon genomic diversity. *Science.* **358**, 951–954 (2017).

26. Lalueza-Fox, C. Museomics. *Curr. Biol.* **32**, R1214–R1215 (2022).

27. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing

Genomic Data Quality and Beyond. *Curr. Protoc.* **1**, e323 (2021).

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

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

30. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).

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

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

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

34. Prüfer, K. SNPAD: An ancient DNA genotype caller. *Bioinformatics* **34**, 4165–4171 (2018).

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

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

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

38. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

39. Allio, R. *et al.* MitoFinder: Efficient automated large-scale extraction of

mitogenomic data in target enrichment phylogenomics. *Mol. Ecol. Resour.* **20**, 892–905 (2020).

40. Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N. & Delsuc, F. MACSE v2: Toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.* **35**, 2582–2584 (2018).

41. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

42. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

43. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

44. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

45. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).

46. Bouckaert, R. R. & Drummond, A. J. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evol. Biol.* **17**, 42 (2017).

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

48. Van Zandt Brower, A. Phylogeny of Heliconius Butterflies Inferred from Mitochondrial DNA Sequences (Lepidoptera: Nymphalidae). *Mol. Phylogenet. Evol.* **3**, 159–174 (1994).

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

50. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 1–8 (2007).

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

52. Keightley, P. D. *et al.* Estimation of the Spontaneous Mutation Rate in Heliconius melpomene. *Mol. Biol. Evol.* **32**, 239–243 (2015).

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

54. Lin, J. *et al.* Probing the genomic limits of de-extinction in the Christmas Island rat. *Curr. Biol.* **32**, 1650-1656.e3 (2022).

55. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

56. Gish, W. & States, D. J. Identification of protein coding regions by database similarity search. *Nat. Genet.* **3**, 266–272 (1993).

57. Bushnell, B. BBMap. (2014).

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

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

Tables

Sample Identifier	Generated Reads	Q25 Unique Mapped Reads	Breadth of Coverage	Average Depth Covered Regions
L002	300,294,248	23,337,751	37.27%	5.105
L003	405,198,060	32,547,820	36.86%	6.78
L004	357,165,438	28,722,185	38.77%	6.55
L005	776,312,378	56,459,037	45.7%	12.42
L006	359,520,168	28,498,720	40.07%	6.18
L007	348,916,870	26,758,356	34.79%	6.21
L008	508,120,156	32,107,192	42,08%	7.422
L009	322,955,384	39,312,617	40.6%	8.02
L011	236,886,534	24,165,282	38.6%	5.40
L012	328,359,669	18,683,738	33.37%	4.29
L013	385,635,644	52,612,937	47.2%	12.3

Table 1. Mapping Statistics of the analysed historical specimens. Mapping statistics of the 4 historical *G. xerces* (L003, L005, L007 and L009) and the 7 historical *G. lydagmus* (L002, L004, L006, L008, L011, L012 and L013) specimens mapped against the *G. alexis* reference genome. Average depth is displayed for the covered regions of each individual.

Figure captions

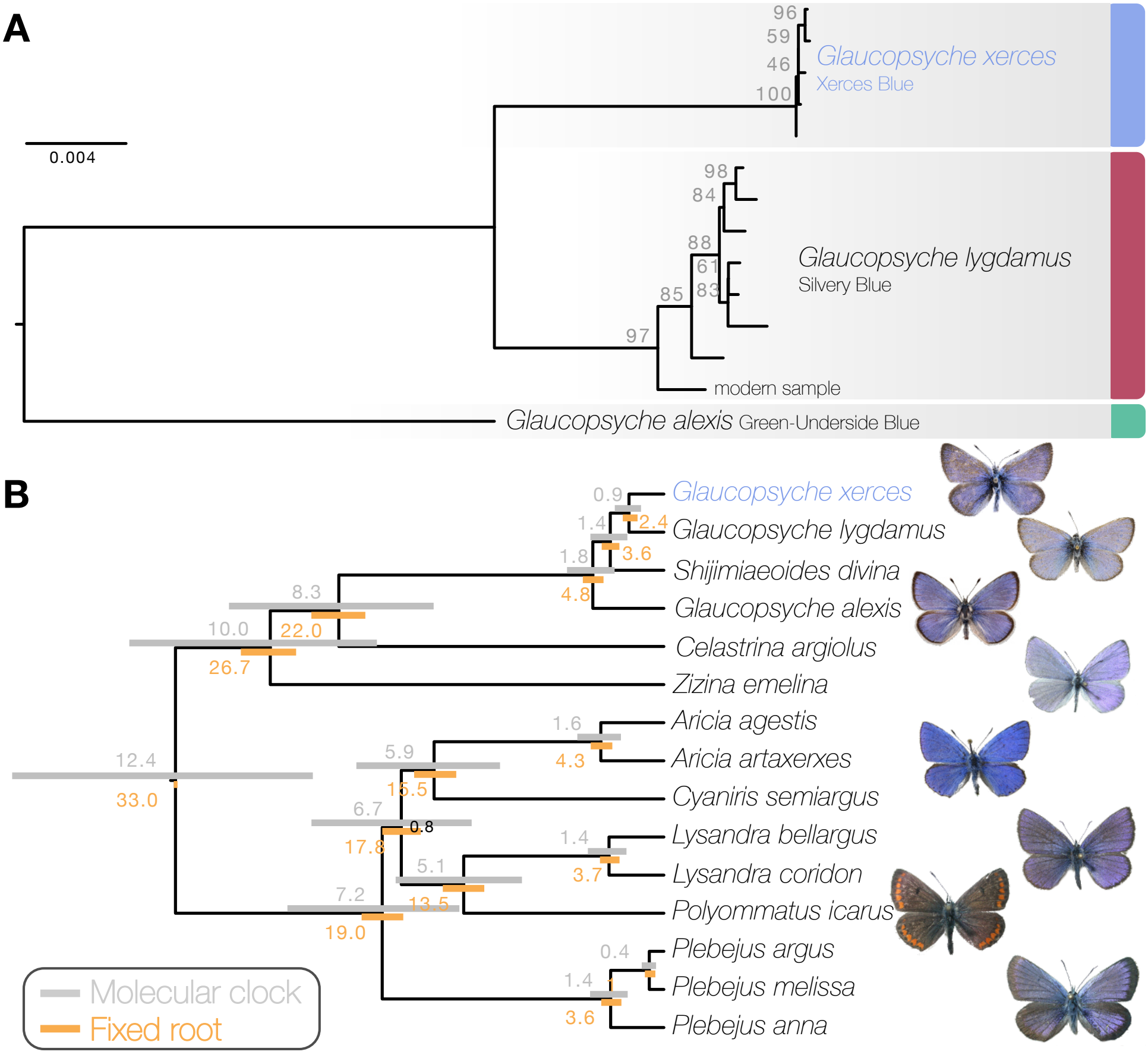
Fig. 1. Phylogenetic placement of the Xerces Blue. A: Maximum likelihood tree from whole mitochondrial genomes of Xerces Blue, Silvery Blue and Green-Underside Blue. Node labels are bootstrap support values. B: Time-calibrated phylogeny from Bayesian inference using mitochondrial protein-coding genes 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. All posterior probabilities were 1, except for one node annotated in black.

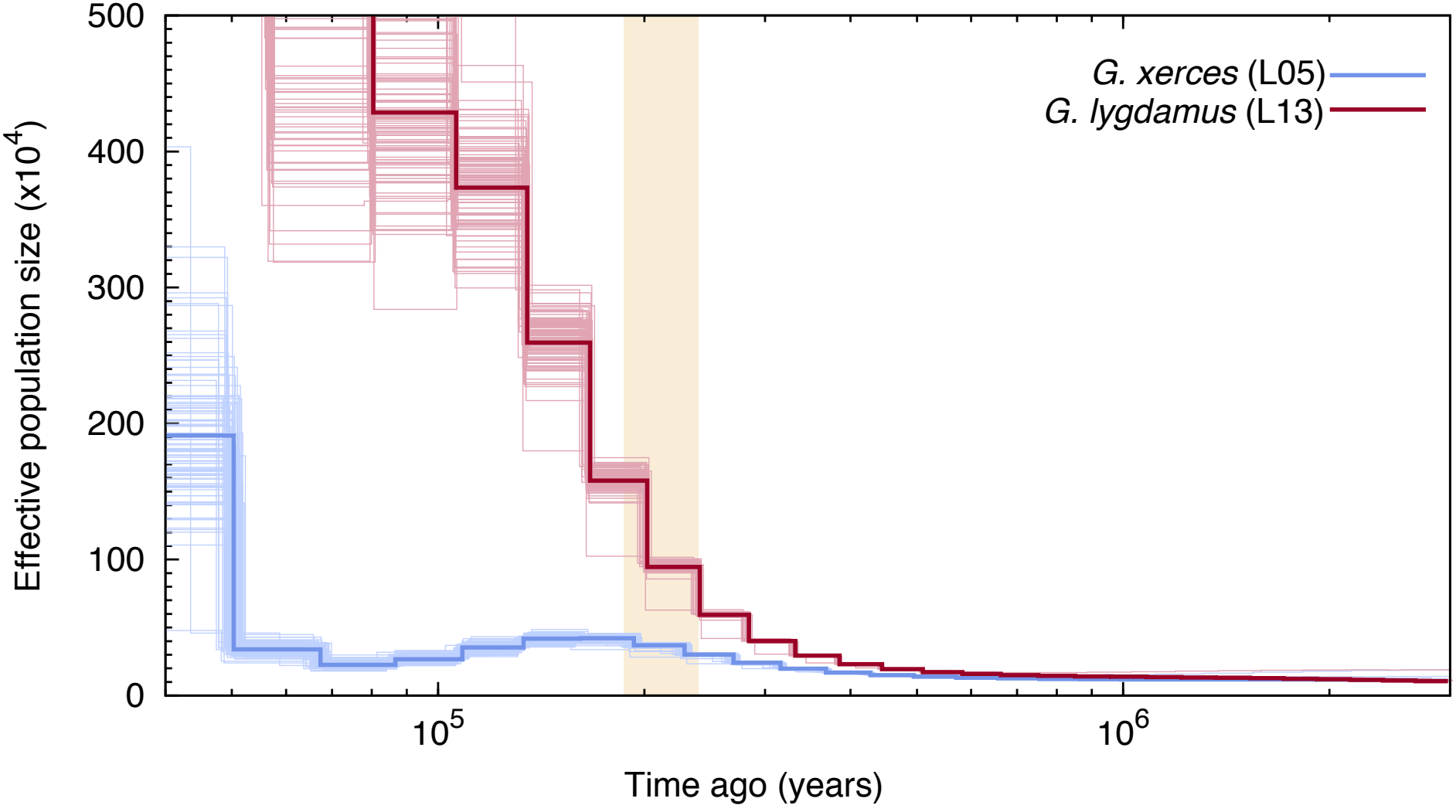
Fig. 2. PSMC plot of one Xerces Blue (*Glaucopsyche xerces*) (L05) specimen and one Silvery Blue specimen (*Glaucopsyche lygdamus*). The two historical samples are those with higher average coverage. Individual PSMC plots were bootstrapped 100 times each (lighter lines). One year of generation time and a mutation rate of $\mu=1.9\times10^{-9}$ were used. The peak of the Marine Isotopic Stage 7 interglacial is marked in yellow.

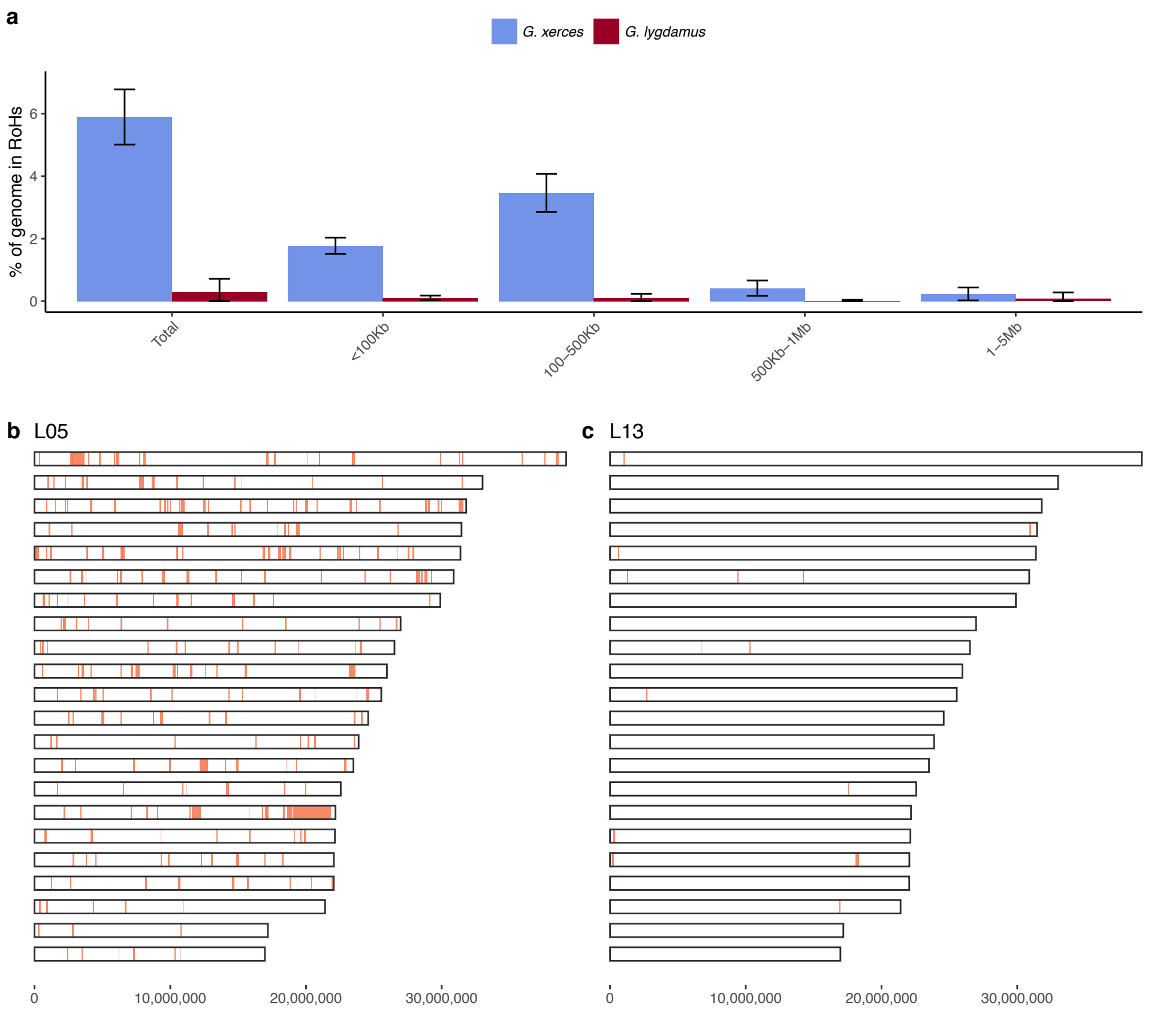
Fig. 3. Runs of Homozygosity (RoH) in the genomes of Xerces Blue and Silvery Blue (modern and historical). a: Percentage of the autosomal genome in RoH by size bins: very short RoH (<100 Kb), short RoH (100-500 Kb), intermediate RoH (500Kb-1Mb) and long (1-5Mb). Short RoH reflect LD patterns, intermediate 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.

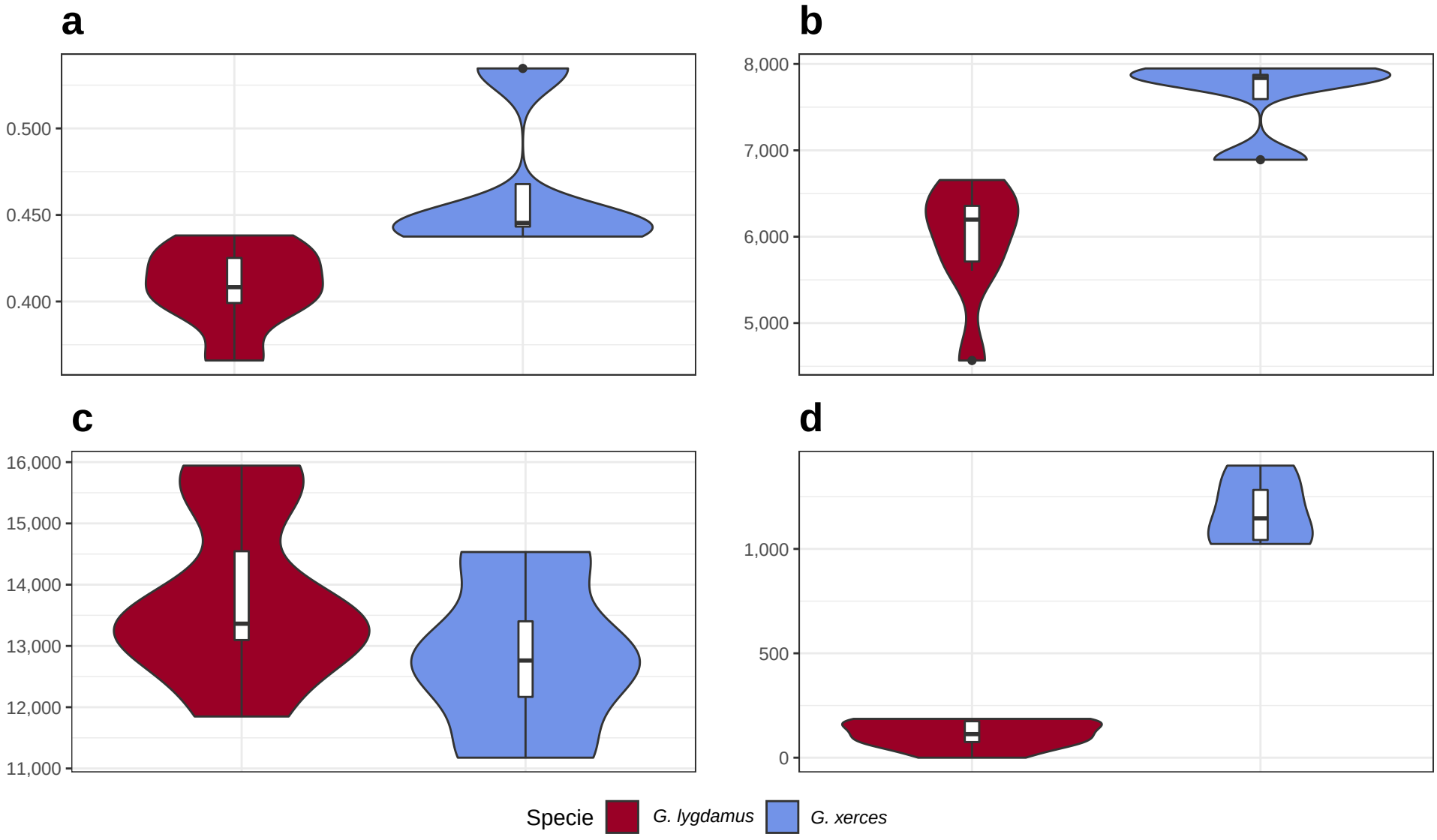
Fig. 4. Functional effect prediction on the fixed amino acid-changing alleles observed in Xerces Blue and Silvery Blue. a: Wide genome Ka/Ks ratio comparison. b: High-to-moderate effect variant comparison in homozygous sites. c: High-to-

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.


A
0.004
96
59
46
100
Glaucopsyche xerces
Xerces Blue
98
84
88
61
83
85
97
Glaucopsyche lygdamus
Silvery Blue
modern sample
Glaucopsyche alexis Green-Underside Blue
B
0.9
1.4
2.4
1.8
3.6
4.8
8.3
22.0
10.0
26.7
12.4
33.0
5.9
15.5
6.7
0.8
17.8
5.1
13.5
7.2
19.0
1.6
4.3
1.4
3.7
0.4
1.4
3.6
Glaucopsyche xerces
Glaucopsyche lygdamus
Shijimiaeoides divina
Glaucopsyche alexis
Celastrina argiolus
Zizina emelina
Aricia agestis
Aricia artaxerxes
Cyaniris semiargus
Lysandra bellargus
Lysandra coridon
Polyommatus icarus
Plebejus argus
Plebejus melissa
Plebejus anna
Molecular clock
Fixed root


500
400
300
200
100
0
Effective population size (x10^4)
10^5
10^6
Time ago (years)
G. xerces (L05)
G. lygdamus (L13)


a
G. xerces
G. lygdamus
% of genome in RoHs
0
2
4
6
Total
<100Kb
100–500Kb
500Kb–1Mb
1–5Mb
b
L05
0
10,000,000
20,000,000
30,000,000
c
L13
0
10,000,000
20,000,000
30,000,000


a
0.500
0.450
0.400
b
8,000
7,000
6,000
5,000
c
16,000
15,000
14,000
13,000
12,000
11,000
d
1,000
500
0
Specie
G. lygdamus
G. xerces