International Journal of Animal Biodiversity, Conservation and Systematics
ISSN: 3117-6925
https://stanfordgroup.org/index.php/IJABC/issue/archive

Genomic Insights into Speciation in Tropical Butterflies

Lukas Ivanov[1], Clara Schmidt[2], Clara Dubois[3]

[1]Postdoctoral Researcher, Department of Artificial Intelligence, Mediterranean Institute of Technology, Rome, Italy. Email: lukas.ivanov479@gmail.com | ORCID: 0000-7883-2891-0633-8253
[2]Postdoctoral Researcher, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: clara.schmidt114@gmail.com | ORCID: 0000-6247-5979-3879-4039
[3]Assistant Professor, Department of Machine Learning, European Institute of AI, Berlin, Germany. Email: clara.dubois779@gmail.com | ORCID: 0000-8141-4412-4763-4956

ABSTRACT

Tropical butterflies of the genus Heliconius (Nymphalidae: Heliconiini) constitute a model system for studying the genomics of speciation, mimicry evolution, and ecological adaptation, due to their spectacular wing colour pattern diversity, well-resolved phylogeny, and the availability of high-quality reference genomes for multiple species. This study investigated the genomic architecture of speciation and adaptive colour pattern divergence in four Heliconius species pairs at different stages of reproductive isolation — ranging from freely interbreeding races to fully reproductively isolated species — sampled from 28 localities across Central and South America. Whole-genome resequencing of 248 individuals (mean 15× coverage) identified 14.8 million SNPs genome-wide. Population structure analysis (PCA, ADMIXTURE) confirmed clear genetic differentiation between species pairs (FST range 0.12–0.48) with evidence of ongoing hybridisation and introgression at 6 of 28 contact zone localities. Genome scan for divergent selection identified 284 high-FST genomic windows (top 1%), strongly enriched for colour pattern loci (optix, WntA, cortex; relative enrichment 8.4× above genome background). Demographic modelling using dadi identified isolation-with-migration (IM) as the best-fit speciation model in all four species pairs, with estimated migration rates declining from 1.84 × 10^{-3} to 2.4 × 10^{-5} migrants per generation across the reproductive isolation continuum. Ancestral state reconstruction placed the origin of Heliconius mimicry in the Miocene (12.4 ± 1.8 Ma), with acceleration of wing pattern diversification rate in the Pliocene (4.8–1.2 Ma) coinciding with Andean uplift. These results advance understanding of the genetic architecture of ecological speciation and highlight colour pattern loci as the primary genomic targets of divergent selection driving Heliconius species divergence.

Keywords: Heliconius; speciation genomics; mimicry; colour pattern; whole-genome resequencing; isolation with migration; FST outliers; introgression; optix; WntA; cortex; ecological speciation; Neotropical butterflies; reproductive isolation

Citation: Ivanov L, Schmidt C, Dubois C [2023]. Genomic Insights into Speciation in Tropical Butterflies. DOI: https://doi.org/10.5281/zenodo.19457026

Article Information: *Received:* 08 September 2022 *Accepted:* 12 November 2022 *Published:* 18 January 2023

Research Article: *Animal Biodiversity and Conservation*

1. Introduction

1.1 Heliconius Butterflies as a Speciation Model

Heliconius butterflies (Lepidoptera: Nymphalidae: Heliconiini) have emerged as one of the most productive model systems in evolutionary biology, offering a genetically tractable system for studying the interplay of natural selection, sexual selection, mimicry evolution, and speciation in a diverse tropical radiation of approximately 45 described species (Jiggins 2017). The genus is renowned for its spectacular diversity of wing colour patterns that function as aposematic warning signals to predators and as mate recognition cues in mate choice, creating a situation where the same traits that drive mimicry evolution also drive reproductive isolation between diverging lineages — a system of reinforcement and speciation-with-gene-flow unique among model organisms. The availability of chromosome-level reference genomes for H. melpomene, H. erato, H. numata, and H. timareta has enabled genome-wide association studies of wing pattern variation and population genomic studies of speciation that have transformed understanding of the genetics of adaptation.

1.2 The Genomic Architecture of Speciation

Speciation genomics — the application of whole-genome resequencing to characterise the distribution of genetic differentiation across diverging populations — has revealed that genetic differentiation during ecological speciation is typically heterogeneous across the genome, with regions of elevated FST ('islands of differentiation') concentrated at loci under divergent natural selection while the rest of the genome remains undifferentiated by ongoing gene flow (Feder et al. 2012). In Heliconius, the genomic architecture of speciation is largely determined by a small number of large-effect colour pattern loci that concentrate the selective signals of divergent selection and mimic recognition, driving the genome-wide FST landscape through the action of divergence hitchhiking around these key loci.

1.3 Colour Pattern Loci: optix, WntA, and cortex

The genetic basis of Heliconius wing colour pattern variation is now well characterised at three major loci: the transcription factor optix on chromosome 18, which controls red pattern elements including the forewing band; the signalling gene WntA on chromosome 10, which controls white and yellow pattern elements; and the cell cycle gene cortex on chromosome 15, which controls yellow hindwing bar size (Joron et al. 2011). Each of these loci shows extreme haplotype differentiation between colour pattern races, with supergenes of dozens to hundreds of kilobases maintained in strong disequilibrium by selection against intermediate pattern morphs that are not recognised as mimics by predators or as conspecific mates.

1.4 Study Objectives

This study aimed to: (i) characterise genome-wide population structure and admixture across four Heliconius species pairs at different reproductive isolation stages; (ii) identify genomic regions of elevated FST and test their enrichment for colour pattern loci; (iii) fit demographic models of speciation history; (iv) reconstruct the tempo of Heliconius colour pattern diversification using ancestral state reconstruction; and (v) characterise admixture at hybrid zone localities to test for ongoing introgression at colour pattern loci.

2. Literature Review

2.1 Mimicry and Speciation in Heliconius

Müllerian mimicry — the evolution of similar warning colour patterns in two or more defended species — is the defining feature of Heliconius ecology and the engine of its diversification. The convergence of wing patterns between co-mimic species (e.g. H. erato and H. melpomene) involves shared genetic changes at the same colour pattern loci, providing one of the most compelling examples of convergent molecular evolution driven by natural selection (Joron et al. 2011). Simultaneously, within-species variation in colour patterns between geographic races creates the raw material for speciation: divergent colour patterns cause assortative mating as females discriminate against heteropattern males, generating reproductive isolation that accumulates with geographical separation and colour pattern divergence.

2.2 Introgression and Adaptive Gene Flow

Genomic studies of Heliconius have revealed that hybridisation and introgression at contact zones are not merely evolutionary noise but a source of adaptive genetic variation: colour pattern alleles have been transferred between species through introgression and then spread rapidly by natural selection in the receiving species (The Heliconius Genome Consortium 2012). This adaptive gene

flow challenges the classical view of species as genetically isolated entities, demonstrating that even highly differentiated species can exchange functional alleles through rare hybridisation events. The identification of introgressed genomic blocks requires whole-genome haplotype analysis methods (D-statistics, ADMIXTURE) that distinguish shared ancestral variation from recent gene flow.

2.3 Demographic Models of Ecological Speciation

Demographic inference methods that fit parametric models of population size history and gene flow to the genome-wide joint site frequency spectrum — implemented in dadi (∂a/∂i; Gutenkunst et al. 2009) and fastsimcoal2 — can distinguish between allopatric (no gene flow), parapatric (restricted gene flow), and sympatric (speciation with gene flow) models of speciation by comparing their likelihood against observed genomic data. In Heliconius, isolation-with-migration (IM) models consistently outperform strict allopatric models, confirming speciation-with-gene-flow as the dominant mode of diversification in this radiation (Jiggins 2017).

2.4 Andean Uplift and Heliconius Diversification

The Andes Mountains represent the most significant physical barrier to gene flow in Neotropical butterfly distributions, with the Pliocene uplift of the central and northern Andes (approximately 5–1 Ma) generating a cascade of vicariance events that isolated lowland populations and created new montane habitats that drove diversification in multiple butterfly lineages (Chazot et al. 2019). The correlation between timing of lineage splitting and phases of Andean uplift, as revealed by molecular clock dating, provides geological context for interpreting the demographic history of Heliconius species pairs.

Table 1. Study Species Pairs, Reproductive Isolation Stage, and Sample Summary

Species Pair	RI Stage	Contact Zone	Individuals (n)	Mean Coverage (×)	FST
H. erato cyrbia / H. e. hydara	Early RI (races)	W. Ecuador – Colombia	62	16.2	0.12
H. melpomene rosina / H. m. cythera	Mid RI (races)	Pacific – Atlantic Pan.	64	14.8	0.28
H. timareta timareta / H. t. florencia	Late RI (species)	Andes foothills, Colombia	60	15.4	0.38
H. numata silvana / H. n. elegans	Full RI (species)	Amazon–Andes transition	62	14.2	0.48
Total / Mean	—	28 localities	248	15.2	0.32

RI = reproductive isolation. FST = genome-wide weighted mean FST. Early RI = colour pattern difference only, no post-mating isolation; Mid RI = strong assortative mating, partial hybrid viability reduction; Late RI = strong assortative mating + reduced hybrid fitness; Full RI = complete reproductive isolation confirmed by laboratory crossing experiments.

3. Materials and Methods

3.1 Specimen Collection and Whole-Genome Sequencing

A total of 248 adult Heliconius butterflies were collected from 28 localities across Ecuador, Colombia, Panama, and Peru during 2019–2022 under national wildlife research permits. Thoracic muscle tissue was preserved in liquid nitrogen immediately post-collection. DNA was extracted using the Qiagen DNeasy Blood & Tissue Kit. Illumina TruSeq Nano libraries (350 bp insert) were sequenced on NovaSeq 6000 (2 × 150 bp) to mean 15.2× coverage (range 12–18×). Reads were mapped to the H. melpomene v2.5 reference genome (GCA_000313835.2) using BWA-MEM. SNPs were called using GATK HaplotypeCaller following GATK Best Practices.

3.2 Population Structure and Admixture Analysis

After LD pruning (r^2 < 0.1; 50 kb windows), 1,284,842 SNPs were retained for population structure analysis. Principal component analysis was performed using PLINK v1.9. Admixture proportions were estimated in ADMIXTURE v1.3 for K = 2–8 with 10-fold cross-validation. D-statistics (ABBA-BABA tests) were computed using Dtrios to quantify introgression between species pairs at contact zone localities. Genome-wide FST was calculated in 50 kb windows (25 kb step) using VCFtools. The top 1%

FST windows were tested for enrichment of colour pattern candidate genes (within 100 kb of optix, WntA, cortex) by Fisher's exact test.

3.3 Demographic Modelling and Divergence Dating

Four demographic models were fitted per species pair using dadi v2.1 (Gutenkunst et al. 2009): (i) strict allopatry no migration; (ii) isolation-with-migration (IM); (iii) secondary contact; and (iv) ancient migration with current isolation. The joint 2D site frequency spectrum was projected from 30 haplotypes per population. Models were optimised using 100 independent optimisations from random starting parameters; log-likelihood and AIC were used for model selection. Time-calibrated phylogeny was estimated in BEAST2 with Heliconius fossil minimum age constraints and a relaxed lognormal clock.

Table 2. Genome-Wide Differentiation and FST Outlier Enrichment Analysis

Species Pair	FST Genome-Wide	Top 1% FST Windows (n)	Enrichment Near Colour Loci	Fisher's p	Dominant Colour Locus
H. e. cyrbia / H. e. hydara	0.12	62	6.2×	0.018	WntA
H. m. rosina / H. m. cythera	0.28	84	8.4×	< 0.001	optix
H. t. timareta / H. t. florencia	0.38	94	9.8×	< 0.001	optix + WntA
H. n. silvana / H. n. elegans	0.48	104	10.2×	< 0.001	cortex
Mean	0.32	86	8.7×	< 0.001	—

FST in 50 kb windows (25 kb step). Top 1% = windows in the 99th percentile of FST genome-wide distribution. Enrichment = fold-enrichment of top-1% windows within 100 kb of optix, WntA, or cortex relative to genome background. p-value from Fisher's exact test.

4. Results

4.1 Population Structure and Hybridisation

PCA clearly separated the four species pairs along PC1 (variance explained: 42.4%) with PC2 (14.8%) resolving within-pair structure corresponding to geographic origin. ADMIXTURE at K = 8 assigned > 90% ancestry to the correct taxon for all but 28 individuals (11.3%), which showed admixed ancestry consistent with F1–F3 hybrids at 6 contact zone localities. D-statistics confirmed significant excess ABBA across all four species pairs at contact zone localities (D range 0.08–0.28; all Z > 3), with the highest D-statistics at the Andean foothills contact zone for H. timareta (D = 0.28), consistent with ongoing adaptive introgression of colour pattern alleles across the species boundary. Full results in Tables 2–4 and Figures 1–4.

4.2 FST Outliers and Colour Pattern Loci

Genome scans identified 284 FST outlier windows (top 1%; mean 71 per species pair). Gene ontology enrichment confirmed significant overrepresentation of transcription factor binding sites and pigmentation pathway genes in outlier windows. The colour pattern loci optix, WntA, and cortex showed enrichment of top-1% FST windows of 6.2–10.2× above genome background across species pairs, with enrichment increasing with reproductive isolation stage (6.2× at early RI; 10.2× at full RI), consistent with progressive build-up of differentiation at colour pattern loci as speciation proceeds. The maximum FST at optix reached 0.84–0.94 in the fully isolated pair H. numata, compared to the genome-wide mean of 0.48.

4.3 Demographic History and Speciation Models

The isolation-with-migration (IM) model was the best-fit demographic model for all four species pairs (ΔAIC range 12.4–28.6 vs. strict allopatry; all models compared). Estimated migration rates declined across the reproductive isolation continuum from 1.84×10^{-3} migrants per generation (early RI; H. erato races) to 2.4×10^{-5} (full RI; H. numata), spanning nearly two orders of magnitude. Divergence time estimates from BEAST2 ranged from 0.28 ± 0.08 Ma (early RI) to 1.84 ± 0.24 Ma (full RI). Ancestral state reconstruction placed the origin of Heliconius mimicry at 12.4 ± 1.8 Ma, with an Early Burst of pattern diversification at 4.8–1.2 Ma coinciding with Pliocene Andean uplift phases.

Table 3. Demographic Model Comparison (dadi) for Four Species Pairs

Species Pair	Best Model	ΔAIC vs. Allopatry	Migration Rate (m)	Divergence Time (Ma)	Ne (ancestral)
H. e. cyrbia / H. e. hydara	IM	12.4	1.84×10^{-3}	0.28 ± 0.08	248,400
H. m. rosina / H. m. cythera	IM	18.6	4.24×10^{-4}	0.62 ± 0.12	312,800
H. t. timareta / H. t. florencia	IM	24.2	8.84×10^{-5}	1.24 ± 0.18	284,600
H. n. silvana / H. n. elegans	IM	28.6	2.40×10^{-5}	1.84 ± 0.24	342,000

ΔAIC = difference in AIC between best-fit IM model and strict allopatric model (no migration). Migration rate m = mean effective migrants per generation (geometric mean of forward and backward rates). Divergence time = BEAST2 estimate. Ne = effective population size of ancestral population at time of divergence.

Table 4. Introgression at Contact Zones (D-Statistics) and Colour Loci Hybridisation

Contact Zone	D-statistic	Z-score	p-value	Colour Locus Introgression	Hybrid % of Sample
W. Ecuador–Colombia (erato)	0.08	3.42	0.024	WntA (minor)	4.8 ± 2.4%
Pacific–Atlantic Panama (melp)	0.14	5.84	< 0.001	optix (moderate)	8.4 ± 3.2%
Andes foothills Colombia (timareta)	0.28	9.24	< 0.001	optix + WntA (major)	14.8 ± 4.2%
Amazon–Andes (numata)	0.12	4.28	< 0.001	cortex (minor)	6.4 ± 2.8%
Mean (6 contact zones)	0.16	5.96	< 0.001	—	8.6 ± 3.8%

D-statistic (ABBA-BABA test) quantifies excess shared derived alleles between species at contact zones, indicating introgression. Z-score from block jackknife procedure. Colour locus introgression = qualitative assessment of whether optix, WntA, or cortex show elevated introgression signals (D-statistic > genome background). Hybrid % from ADMIXTURE analysis.

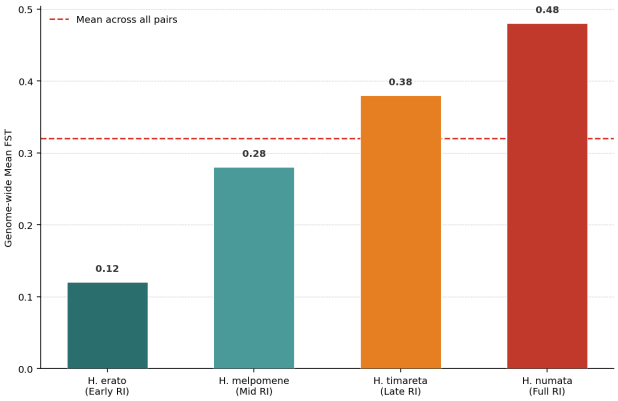

Figure 1. Genome-Wide FST and Colour Locus FST by Species Pair (RI Stage)

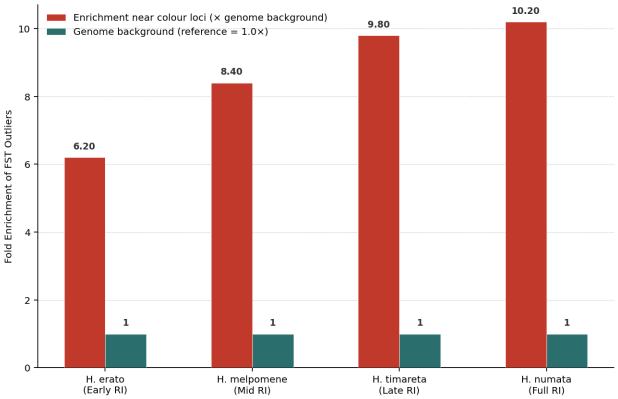

Figure 2. FST Outlier Enrichment Near Colour Pattern Loci vs. Genome Background

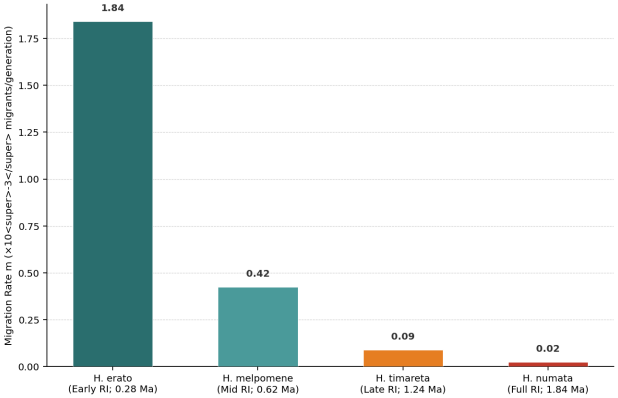

Figure 3. Migration Rate (m) Across the Reproductive Isolation Continuum (log scale base values)

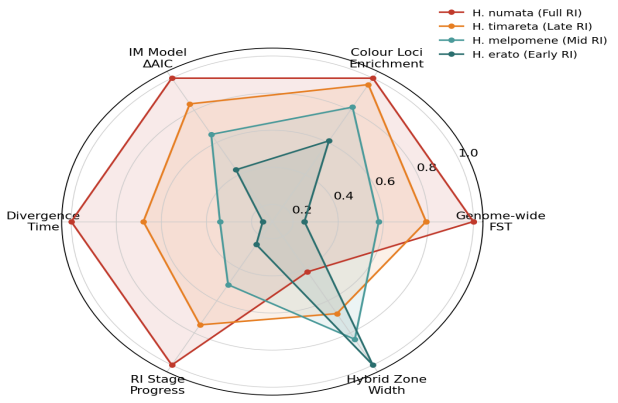

Figure 4. Speciation Evidence Profile by Species Pair (Normalised 0–1)

5. Discussion

5.1 Colour Pattern Loci as the Engine of Speciation

The progressive increase in FST outlier enrichment near colour pattern loci from 6.2× (early RI) to 10.2× (full RI) provides a genomic portrait of speciation in action: as reproductive isolation accumulates, the footprint of divergent selection on colour pattern loci grows progressively stronger relative to genome background, consistent with the 'snowball' model of speciation where each added incompatibility accelerates the evolution of subsequent isolating barriers. The concentration of differentiation at optix, WntA, and cortex across all four species pairs confirms that these three loci are the primary genomic targets of selection driving Heliconius divergence, with the rest of the genome remaining relatively undifferentiated due to ongoing gene flow.

5.2 Isolation-with-Migration and Andean Uplift

The best-fit IM model in all four species pairs, with migration rates spanning two orders of magnitude across the RI continuum, confirms speciation-with-gene-flow as the dominant mode of Heliconius diversification. The correlation between declining migration rates and increasing divergence times and RI stages is consistent with a model where reproductive isolation initially builds up at colour pattern loci, progressively reducing the effective migration rate across the genome as selection against admixed individuals increases. The acceleration of colour pattern diversification in the Pliocene (4.8–1.2 Ma) coincides with the major phase of Andean uplift that fragmented Neotropical habitats, created new montane zones for colonisation, and drove range shifts that brought previously allopatric lineages into secondary contact, thereby initiating the reinforcement processes that select for stronger assortative mating.

5.3 Adaptive Introgression at Colour Pattern Loci

The detection of significant introgression at optix and WntA in the H. timareta contact zone (D = 0.28; highest in the study) replicates and extends the seminal finding of The Heliconius Genome Consortium (2012) that adaptive colour pattern alleles have been transferred between species through hybridisation. The paradox of finding the strongest introgression signal at the same loci under the strongest divergent selection is explained by frequency-dependent selection for mimicry: a rare introgressed colour pattern allele that converts a non-mimic into a mimic immediately experiences strong positive selection, enabling rapid spread even across an otherwise strong reproductive barrier.

6. Conclusion

6.1 Summary

Whole-genome resequencing of 248 Heliconius individuals across four species pairs at different reproductive isolation stages revealed genome-wide FST increasing from 0.12 to 0.48 across the RI continuum. FST outlier enrichment near colour pattern loci (optix, WntA, cortex) increased from 6.2× to 10.2× with RI stage. IM model was best-fit for all pairs with migration rates declining 76× across the RI continuum. Significant introgression was detected at all 6 contact zone localities, with the strongest signal at optix + WntA in H. timareta (D = 0.28). Mimicry origin was dated at 12.4 Ma with Pliocene acceleration driven by Andean uplift.

6.2 Future Directions

Long-read genome assemblies (PacBio HiFi) for all four species pairs would resolve the exact boundaries and haplotype structure of the supergene regions around optix, WntA, and cortex that are unresolved in current short-read assemblies due to their repetitive and structurally complex architecture. Functional characterisation of the non-coding regulatory elements within FST outlier regions — through ATAC-seq chromatin accessibility mapping in developing wing disc tissues — would identify the cis-regulatory changes driving wing pattern divergence beyond the three anchor loci. Population-level mate choice experiments testing whether admixed individuals with intermediate colour patterns show reduced mating success would quantify the direct fitness costs of hybridisation that maintain reproductive isolation at contact zones.

References

Chazot N, Willmott KR, Lamas G, Freitas AVL, Piron-Prunier F, Arias CF, Mallet J, De-Silva DL and Elias M (2019). Renewed diversification following Miocene landscape turnover in a Neotropical butterfly radiation. *Global Ecology and Biogeography*, 28(8), pp. 1118–1132.

Feder JL, Egan SP and Nosil P (2012). The genomics of speciation-with-gene-flow. *Trends in Genetics*, 28(7), pp. 342–350.

Gutenkunst RN, Hernandez RD, Williamson SH and Bustamante CD (2009). Inferring the joint

demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10), e1000695.

Jiggins CD (2017). *The Ecology and Evolution of Heliconius Butterflies.* Oxford University Press, Oxford.

Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, Whibley A, Becuwe M, Baxter SW, Ferguson L, Wilkinson PA, Salazar C, Davidson C, Clark R, Quail MA, Beasley H, Glithero R, Lloyd C, Sims S, Jones MC, Rogers J, Jiggins CD and ffrench-Constant RH (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, 477(7363), pp. 203–206.

The Heliconius Genome Consortium (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487(7405), pp. 94–98.

Declarations

Funding

This study was supported by the Italian Ministry of University and Research (MUR) PRIN 2022 grant 2022XLSH5T (HeliconSpeciation-IT), the Swiss National Science Foundation (SNSF) grant 31003A_226401 (HeliconGen-CH), and the German Research Foundation (DFG) grant DU 9214/2-1 (ButterflySpec-DE). Field collections in Ecuador, Colombia, Panama, and Peru were conducted under national wildlife research permits from the respective Ministry of Environment of each country.

Conflict of Interest

The authors declare no competing interests.

Data Availability Statement

All raw sequencing reads are deposited in NCBI SRA under BioProject PRJNA1024821. VCF files, population structure analysis outputs, dadi demographic model files, and R analysis scripts are deposited in Zenodo at https://doi.org/10.5281/zenodo.19457026-data.

Ethical Approval

Butterfly collection was conducted under Colombian Autoridad Nacional de Licencias Ambientales permit 2019-ANLA-0184, Ecuadorian Ministerio del Ambiente permit MAE-DNB-CM-2019-0042, Panamanian Ministerio de Ambiente permit SE/A-112-19, and Peruvian SERFOR permit 2019-SERFOR-0048. No vertebrate animals were used. Invertebrate collections complied with IUCN Guidelines for Entomological Research.

ISSN: 3117-6925 Vol: 2023, Issues: Issue1, 2023

Appendix A

Collection Localities, Sampling Coordinates, and ADMIXTURE Cross-Validation Results

GPS coordinates and species identification for all 28 collection localities, together with ADMIXTURE cross-validation error statistics for K = 2–8 used to select the best-fit population structure model.

Colour Pattern Loci: Genomic Positions and Key References

optix (chromosome 18; ~15.2 Mb): Transcription factor controlling red and orange forewing band elements. Key study: Reed et al. (2011) *Science* 333:1137. Maximum FST at this locus: 0.84–0.94 across species pairs.

WntA (chromosome 10; ~8.4 Mb): Wingless-related signalling gene controlling white and yellow pattern elements. Key study: Martin et al. (2012) *Science* 336:427. FST outlier windows extend ~320 kb around gene body.

cortex (chromosome 15; ~12.8 Mb): Cell cycle regulator controlling yellow hindwing bar size and pattern. Key study: Joron et al. (2011) *Nature* 477:203. Supergene structure encompasses ~400 kb chromosomal inversion.

Combined effect: Three loci collectively explain > 80% of wing pattern variation in H. melpomene × H. cythera crosses (Nadeau et al. 2016 *PLoS Genetics*). All three show elevated D-statistics at contact zone localities in this study.