Supergene formation is associated with a major shift in genome-wide patterns of diversity in a butterfly

María Ángeles Rodríguez de Cara[1*$], Paul Jay[1*$], Mathieu Chouteau[1,2], Annabel Whibley[3,4], Barbara Huber[5], Florence Piron-Prunier[3], Renato Rogner Ramos[6], André V. L. Freitas[6], Camilo Salazar[7], Karina Lucas Silva-Brandão[8], Tatiana Texeira Torres[9], Mathieu Joron[1$]

* contributed equally

[1]Centre d'Ecologie Fonctionnelle et Evolutive (CEFE), Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France

[2]Laboratoire Ecologie, Evolution, Interactions Des Systèmes Amazoniens (LEEISA), Université de Guyane, IFREMER, CNRS, Cayenne, Guyane Française

[3]Institut de Systématique Evolution Biodiversité (ISYEB), Museum National d'Histoire Naturelle, CNRS, Sorbonne-Université, EPHE, Université des Antilles, Paris, France

[4]School of Biological Sciences, University of Auckland, Auckland, New Zealand

[5]Instituto de Ciencias Ecológicas y Ambientales (ICAE), Univ de los Andes, Mérida, Venezuela

[6]Departamento de Biologia Animal, Instituto de Biologia, Unicamp, Campinas, São Paulo, Brazil

[7]Department of Biology, Faculty of Natural Sciences, Universidad del Rosario, Carrera 24 No 63C-69, Bogotá 111221, Colombia.

[8]Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas. Av. Candido Rondom 400. Campinas, São Paulo, Brazil

[9]Department of Genetics and Evolutionary Biology, Institute of Biosciences, University of São Paulo (USP), São Paulo, Brazil

$ Corresponding authors: angeles.decara@gmail.com, paul.yann.jay@gmail.com, mathieu.joron@cefe.cnrs.fr

Abstract: Selection shapes genetic diversity around target mutations, yet little is known about how selection on specific loci affects the genetic trajectories of populations, including their genome-wide patterns of diversity and demographic responses. Adaptive introgression provides a way to assess how adaptive evolution at one locus impacts whole-genome biology. Here we study the patterns of genetic variation and geographic structure in a neotropical butterfly, *Heliconius numata*, and its closely related allies in the so-called melpomene-silvaniform subclade. *H. numata* is known to have evolved a supergene via the introgression of an adaptive inversion about 2.2 million years ago, triggering a polymorphism maintained by balancing selection. This locus controls variation in wing patterns involved in mimicry associations with distinct groups of co-mimics, and butterflies show disassortative mate preferences and heterozygote advantage at this locus. We contrasted patterns of genetic diversity and structure 1) among extant polymorphic and monomorphic populations of *H. numata*, 2) between *H. numata* and its close relatives, and 3) between ancestral lineages in a phylogenetic framework. We show that *H. numata* populations which carry the introgressed inversions in a balanced polymorphism show markedly distinct patterns of diversity compared to all other taxa. They show the highest diversity and demographic estimates in the entire clade, as well as a remarkably low level of geographic structure and isolation by distance across the entire Amazon basin. By contrast, monomorphic populations of *H. numata* as well as its sister species and their ancestral lineages all show the lowest effective population sizes and genetic diversity in the clade, and higher levels of geographical structure across the continent. This suggests that the large effective population size of polymorphic populations could be a property associated with harbouring the supergene. Our results are consistent with the hypothesis that the adaptive introgression of the inversion triggered a shift from directional to balancing selection and a change in gene flow due to disassortative mating, causing a general increase in genetic diversity and the homogenisation of genomes at the continental scale.

Introduction: Genetic diversity is shaped by selective processes such as stabilizing or disruptive selection, and by demographic processes such as fluctuations in effective population size. Empirical studies on genetic diversity within and among populations abound, fuelled by an increasing availability of whole genome data, and spurred by our interest in understanding the underlying causes of variation in diversity (e.g. Beichmann 2018, Muers 2009; Murray 2017; Nielsen et al. 2009). At the locus scale, strong directional or disruptive selection tends to reduce diversity within populations (Mitchell-Olds et al. 2007), while balancing selection tends to enhance diversity (Charlesworth 2006). Genome-wide factors reducing diversity include low effective population sizes, generating drift, while high genetic diversity is enhanced by large population sizes and gene flow. Overall, it is well recognised that demographic changes should have a genome-wide effect on diversity, while positive selection is expected to play a role on the sites within and around the genes involved in trait variation (Glinka et al. 2003, Muers 2009, Nielsen et al. 2009).

Variation in behaviour and life-history traits, for instance involving changes in offspring viability or dispersal distance, may also affect species demography, and thus whole genome genetic diversity. However, whether and how genetic variability in a population may be driven by phenotypic evolution at certain traits is poorly understood, and confounding effects may affect patterns of genomic diversity, such as variation in census population size or colonization history. Dissecting how selection on a trait may affect genome-wide diversity can be tackled by comparing closely-related populations differing at this trait coupled with knowledge of when the differences evolved. Here, we took advantage of the dated introgressive origin of a chromosomal inversion associated

with major life-history variation to study the demographics and whole genome consequences of changes in the selection regime at a major-effect locus.

Heliconius butterflies are aposematic, chemically-defended butterflies distributed over the American tropics from Southern Brazil to Southern USA (Emsley 1965; Brown 1979) (Fig 1A). *Heliconius* butterflies are well-known for visual resemblance among coexisting species, a relationship called Müllerian mimicry which confers increased protection from bird predators through the evolution of similar warning signals (Sheppard et al. 1985). Most species are locally monomorphic, but their mimicry associations vary among regions, and most species display a geographic mosaic of distinct mimetic "races" through their range. In contrast to most *Heliconius* species, the tiger-patterned *Heliconius numata* is well-known for maintaining both mimicry polymorphism within localities, with up to seven differentiated coexisting forms, and extensive geographic variation in the distribution of wing phenotypes (Brown & Benson 1974; Joron et al. 1999). Forms of *H. numata* combine multiple wing characters conveying resemblance to distinct sympatric species in the genus *Melinaea* and other local Ithomiini species (Nymphalidae: Danainae). Polymorphism in *H. numata* is controlled by a supergene, i.e. a group of multiple linked functional loci segregating together as a single Mendelian locus, coordinating the variation of distinct elements of phenotype (Brown & Benson 1974; Joron et al. 2006). Supergene alleles are characterized by rearrangements of the ancestral chromosomal structure, forming three distinct chromosomal forms with zero (ancestral type, Hn0), one (Hn1) or three chromosomal rearrangements (Hn123) (Fig 1B). The ancestral arrangement, Hn0, devoid of inversions, is fixed in most *Heliconius* species (although an inversion in the same region evolved independently in a distantly-related *Heliconius* lineage (Edelman et al. 2019)). Arrangement Hn1 contains a 400kb inversion called P_1 originating from an introgression event about 2.2 My ago from *H. pardalinus*, in which P_1 is fixed (Jay et al. 2018). This introgression is thought to be the founding event triggering the formation of the supergene and the maintenance of polymorphism in *H. numata* (Jay et al. 2018). Arrangement Hn123 displays two additional inversions, P_2 and P_3, in linkage with P_1, and therefore originated after the introgression of P_1 into the *H. numata* lineage (Jay et al. 2021).

Heliconius numata is widespread in the lowland and foothill tropical forests of the Amazon basin, the Guianas, and the Brazilian Atlantic Forest (Mata Atlântica), but the frequencies of the three chromosome arrangements vary across the range. Ancestral type Hn0 is fixed in the Atlantic Forest populations of Brazil (forms *robigus* or *ethra)*, but segregates at intermediate frequencies in all other *H. numata* populations throughout the range (forms *silvana* and *laura*) (Fig 1C). Chromosome type Hn1 is associated with the Andean mimetic form *bicoloratus* and is found in the Eastern Andean foothills of Ecuador, Peru, and Bolivia. Chromosome type Hn123 is associated with a large diversity of wing-pattern forms of intermediate allelic dominance, including *tarapotensis, arcuella* and *aurora*, and is reported from Andean, lowland Amazonian and Guianese populations. Inversion polymorphism is therefore structured across the range, with populations being fixed for the ancestral chromosome (Atlantic Forest, see Text S1 & Table S1-2), or displaying a polymorphism with two (Amazon-Guiana) or three (Andes) chromosomal types in coexistence (Joron et al. 2011). Monomorphic populations of the Atlantic forest, devoid of rearrangements at the supergene locus, might represent the ancestral state displayed by *H. numata* populations before the evolution of the supergene via introgression (Fig 1C).

The wing patterns of *H. numata* are subject to selection on their resemblance to local co-mimics

(Chouteau et al. 2016), but the polymorphism is maintained by balancing selection on the chromosome types. Balancing selection is indeed mediated by disassortative mating favouring mixed-form mating (Chouteau et al. 2017) and is likely to have evolved in the response to the deleterious mutational load carried by inversions, which causes heterozygous advantage in *H. numata* (Jay et al. 2021, Faria et al. 2019, Maisonneuve et al. 2019). The introgression of P_1 and the formation of a supergene were associated with a major shift in the selection regime and in the mating system and may therefore have profoundly affected the population biology of the recipient species, *H. numata*. We investigate here whether the adaptive introgression of a balanced inversion is associated with a signature in the genetic diversity and geographic structure. We analyse changes in the demographic history of the clade containing *H. numata* and closely related taxa, as well as their current patterns of diversity and demography, using three well separated populations of *H. numata* representing different states of inversion polymorphism. Our results are consistent with the selection regime and mating system associated with supergene formation having enhanced gene flow among populations and increased effective population size. Moreover, our findings highlight that balancing selection and a shift in mating systems associated with chromosomal polymorphism may reshape genomewide diversity, with crucial consequences on current patterns of genetic structure and population ecology.

Material and Methods

We used here whole genome resequencing from 137 specimens of *Heliconius*, including 68 *H. numata*. Sampling included specimens from populations in the Andean foothills (3 chromosome types), from the upper Amazon (2 chromosome types), from French Guiana (2 chromosome types) and from the Brazilian Atlantic Forests (1 chromosome type) (Fig 1C; Table S3). Related taxa were represented by the sister species *H. ismenius*, found west of the Andes (parapatric to *H. numata*), by Amazonian representatives of the lineage *H. pardalinus* (donor of the inversion), *H. elevatus*, *H. ethilla*, *H. besckei* as well as *H. hecale*, and by *H. melpomene* and *H. cydno* as outgroups. Only Andean, Amazonian and Guianese populations of *H. numata* display chromosomal polymorphism, all other taxa being fixed for the standard gene arrangement (Hn0), or for the inverted arrangement Hn1 (*H. pardalinus*) (Jay et al. 2018). Hereafter, *H. numata* populations from the Andes, Amazon and French Guiana will be collectively referred to as “Amazonian”, and populations from the Atlantic Forest as “Atlantic”. Butterfly bodies were preserved in NaCl saturated DMSO solution at 20°C and DNA was extracted using QIAGEN DNeasy blood and tissue kits according to the manufacturer’s instructions with RNase treatment. Illumina Truseq paired-end whole genome libraries were prepared and 2x100bp reads were sequenced on the Illumina HiSeq 2000 platform. Reads were mapped to the *H. melpomene* Hmel2 reference genome (Davey et al., 2016) using Stampy (version 1.0.28; Lunter and Goodson, 2011) with default settings except for the substitution rate which was set to 0.05 to allow for the expected divergence from the reference of individuals in the so-called silvaniform clade (*H. numata*, *H. pardalinus*, *H. elevatus*, *H. hecale*, *H. ismenius*, *H. besckei* and *H. ethilla*). *H. melpomene* and *H. cydno* belonging to the so-called *melpomene* clade, their genomes were mapped with a substitution rate of 0.02. Alignment file manipulations were performed using SAMtools v0.1.3 (Li et al. 2009). After mapping, duplicate reads were excluded using the *MarkDuplicates* tool in Picard (v1.1125; http://broadinstitute.github.io/picard) and local indel realignment using IndelRealigner was performed with GATK (v3.5; DePristo et al. 2011). Invariant and polymorphic sites were called with GATK HaplotypeCaller, with options --min_base_quality_score 25 --min_mapping_quality_score 25 -stand_emit_conf 20 --heterozygosity 0.015.

F_{ST}, d_{XY} and π, were calculated in overlapping windows of 25 kb based on linkage disequilibrium decay (*Heliconius* Genome Consortium 2012) using custom scripts provided by Simon H. Martin (https://github.com/simonhmartin), and the genome-wide average was calculated using our own scripts (available from https://github.com/angelesdecara). Distance in km between sampling sites was measured along a straight line, not taking into account potential physical barriers. The slopes of F_{ST} versus distance was calculated using the R package *lsmeans* (Lenth 2016); the slope difference among species or between populations within species was estimated with an ANOVA and its significance evaluated with function pairs of this package (Text S1 and see example script on github.com/angelesdecara/).

Admixture (Alexander et al. 2009) analyses were run on a subset of the 68 *H. numata* genomes, keeping only 15 individuals from Peru to have a more balanced representation of individuals across the geographic distribution. Filters were applied to keep biallelic sites with minimum mean depth of 8, maximum mean depth of 200 and at most 50% genotypes missing. We only kept 1 SNP per kilobase to remove linked variants, and we obtained the optimal number of clusters using cross-validation for values of K from 1 to 10 (Alexander et al. 2009). Principal component analyses (PCA) were performed with the same filters as for admixture, using the same *H. numata* genomes as for the admixture analyses, using smartpca (Patterson et al. 2006).

In order to estimate demographic parameters independently of the effect of selection on diversity, we performed stringent filtering on the dataset. We removed all predicted genes and their 10,000 base-pair flanking regions, before performing G-PhoCS (Gronau et al. 2011) analyses as detailed below. Repetitive regions were masked using RepeatMasker and Tandem Repeat Finder (Benson 1999). GC islands detected with CpGcluster.pl with parameters 50 and 1E-5 (Hackenberg et al., 2006) were also masked. Scaffolds carrying the supergene rearrangements (Hmel215006 to Hmel215028) were excluded, as were scaffolds from the sex chromosome (Z), since those are expected to show unusual patterns of diversity due to selection and different effective population sizes.

We analysed the demographic history of *H. cydno*, *H. numata*, *H. ismenius*, *H. pardalinus* and *H. elevatus* with G-PhoCS, which allows for the joint inference of divergence time, effective population sizes and gene flow. In order to detect differences in demography correlating with the presence of the supergene in *H. numata*, we conducted analyses separating the Atlantic population of *H. numata* from Amazonian populations. G-PhoCS is an inference method based on a full coalescent isolation-with-migration model. Inferences are conditioned on a given population phylogeny with migration bands that describe allowed scenarios of post-divergence gene flow. The model assumes distinct migration rate parameters associated with each pair of populations, and allows for asymmetric gene flow. Given the computational burden of G-PhoCS, we selected two individuals per taxon or population, retaining those with the highest sequencing depth (see Table S3). The input dataset consisted of 4092 genomic regions, each 1kb in length and spaced at approximately 30kb intervals and with genotypes in at least one of the two samples of each taxon We used as priors for coalescence times (τ) and genetic diversity (θ), Gamma functions with α=1 and β=100, and for migration bands α=0.002 and β=0.00001. These priors were chosen to allow good convergence while also ensuring non informativity. In order to calculate the highest posterior density interval, we used the library HDInterval in R, and to integrate such posterior densities we

used the library sfsmisc in R. We rescaled the results using a mutation rate of 1.9E-9 (Martin et al. 2016) and 4 generations per year (i.e., g=0.25). Migration bands were considered significant following the criteria of Freedman et al. (2012): if the 95% HPD interval did not include 0 or if the total migration was larger than 0.03 with posterior probability larger than 0.5.

Results

Using cross validation error as a measure of the optimal number of clusters with Admixture, we found that K=2 was the optimal cluster number describing within-species genetic variation in *H. numata* (Fig 2A). One cluster corresponds to the Atlantic population, forming a well-differentiated genetic entity compared to all other *H. numata* populations. All Amazonian populations of *H. numata* showed a remarkable uniformity, with the exception of a few individuals sharing some variation with SE Brazil. This pattern is consistent with the population structure inferred using microsatellite markers (Fig S1). Population structure revealed by PCA is in line with the admixture analysis (Fig 2B). Individuals from the Atlantic population of *H. numata* clustered together to one side of the first PCA axis, whereas all other individuals from all other populations clustered to the other side. The second axis of the PCA separates individuals from French Guiana from the other samples of the Amazon. This clustering was not found with Admixture (i.e. with K=3), suggesting that the divergence between Amazonian populations is very reduced. In accordance, pairwise genome-wide estimates of differentiation (F_{ST}) between *H. numata* populations showed elevated values when comparing the Atlantic population to other populations, but very small values when comparing pairs of Amazonian populations, even at a large distance (Fig 2C, Table S4). For instance, the population from La Merced in Peru shows an F_{ST}=0.032 with the population from French Guiana at a distance of 3019km, but an F_{ST}=0.311 (an order of magnitude higher) with the Atlantic population at a similar distance. Isolation by distance among Amazonian populations of *H. numata*, estimated using the proxy F_{ST}/km, shows a very different pattern to other species, with a highly significantly shallower increase in F_{ST} with distance in *H. numata* compared to all other taxa (Fig 2C, Table S4). By contrast, differentiation as a function of distance between Atlantic and Amazonian populations of *H. numata* is close to what is observed in other species, and not significantly different (see Supp. Text S1).

Analyses of genetic diversity show that all populations of *H. numata*, except those from the Atlantic Forest, have a similar high genetic diversity (Fig 3A). By comparison, closely related *Heliconius* taxa show significantly lower genetic diversity (Fig 3A). These patterns are similar to those obtained using G-PhoCS to analyse the demographic histories in a phylogenetic context, where Amazonian populations of *H. numata* show higher population sizes compared to the Atlantic population (Fig 3B, Table S5). G-PhoCS analyses also show a demographic history in which gene flow plays a crucial role (Table S6). For instance, our analyses show strong significant gene flow right at the beginning of the divergence between *H. ismenius* and the other silvaniforms, as well as in the divergence between *H. pardalinus* and *H. elevatus.* The effective population sizes inferred from Atlantic genomes are one order of magnitude lower than that obtained using *H. numata* populations from other localities (Fig 3A and Table S5). In our cladogram, the increase in *H. numata* population size is restricted to the Amazonian branch, excluding Atlantic populations.

Discussion

Our results suggest that populations displaying inversion polymorphism in the *P* supergene in *H. numata* also display distinctive population demography and gene flow. Differences in demographic and differentiation regimes associated with structural variation at this locus are revealed when comparing polymorphic populations of *H. numata* to closely-related monomorphic taxa, such as (1) peripheral populations of *H. numata*, (2) sister taxa, and (3) inferred ancestral lineages. This suggests that the existence of a mimicry supergene controlling polymorphism in *H. numata* is associated, in time and in space, with major differences in population biology. We hypothesize this to be due to a change in the balancing selection regime due to heterozygote advantage (Jay et al. 2021) and in the associated evolution of disassortative mating (Chouteau et al. 2017) following the onset of inversion polymorphism, causing direct effects on ecological parameters such as gene flow, immigration success and effective population size.

Our analyses show large-scale variation in genetic diversity among closely related taxa in this clade of *Heliconius* butterflies. Within *H. numata*, the genetic diversity of polymorphic Amazonian populations is one to two orders of magnitude higher than the diversity found in populations from the Atlantic Forest. Generally, Amazonian populations of *H. numata* harbour the highest genetic diversity in the entire *melpomene*/silvaniform clade, which contrasts with the low diversity found in the most closely related taxa such as *H. ismenius* or *H. besckei*. Inferring historical demography during the diversification of the *H. numata* lineage reveals that the large effective population size in that species is only associated with the branch representing polymorphic, Amazonian *H. numata* populations, while internal branches all show very low diversity estimates. This suggests that ancestral monomorphic populations of *H. numata* were similar in their diversity parameters to current sister species *H. ismenius* populations, or to current peripheral Atlantic *H. numata* populations. Although low-diversity lineages could have lost diversity due to recent events such as strong bottlenecks, the distribution of parameters across lineages rather suggests that the Amazonian populations of *H. numata* underwent a dramatic increase in effective population size posterior to their split with Central American (*H. ismenius*) and Atlantic populations. The Amazonian branch of the *H. numata* radiation is characterized by the long-term maintenance of inversion polymorphism, triggered by the introgression of a chromosomal inversion about 2.2 Ma ago. Therefore, the major shift in demography between Amazonian and Atlantic populations indeed appears associated with the occurrence of inversion polymorphism, even though the lack of replication of this event impedes firmly establishing causality here.

Another striking result is the low genetic structure displayed by *H. numata* across the Amazon, with all Amazonian and Guianese populations forming a single genetic cluster. Only Atlantic populations stand out and display high differentiation with other *H. numata* from the rest of the range. French Guiana and Peruvian populations, separated by over 3000 km across the Amazon, are remarkably genetically similar compared to pairs of populations at comparable distances in other species, and show similar differentiation as pairs of *H. numata* populations taken at short distances. *H. numata* populations from the Amazon show significantly lower isolation by distance than all other taxa, as measured by the change in F_{ST} across distance (F_{ST}/km) (Fig. 2C), with a very distinctive, flat slope of isolation by distance. The only exception is found when comparing Amazonian populations with Atlantic populations of Brazil, displaying a level of differentiation in line with that of pairs of populations at similar distances within other taxa.

Effective population size is affected by census size, mating system, and the force and type of selection acting on traits (Charlesworth 2009). Selection is often viewed as a force only affecting the genetic variation around specific, functional loci in the genome, but it may also affect whole genome diversity, for instance when its action is sufficient to modify local demography or mating patterns. In *H. numata*, morphs and therefore inversion genotypes show disassortative mate preferences, i.e., they preferentially mate with individuals carrying different chromosome types (Chouteau et al. 2017). Disassortative mating enhances heterozygosity and the mating success of individuals expressing rare alleles (negative frequency dependence) (Knoppien 1985; Hedrick et al. 2018). Consequently, immigrants expressing rare, recessive alleles have a mating advantage in *H. numata.* Their recessive effect on wing pattern lets them escape negative selection caused by their inadequate mimicry patterns. Disassortative mating associated with the supergene should therefore bring an advantage to immigrant genomes in LD with recessive supergene alleles, enhancing genome-wide gene flow. This effective migration regime is quite different to that observed in other mimetic taxa such as *Heliconius melpomene* or *H. erato,* in which mimicry variation is controlled by multiple loci with diverse dominance patterns. In those taxa, hybrid offspring display recombinant patterns breaking down mimicry, even after multiple generations of backcrossing, and pure forms mate assortatively with respect to wing pattern (McMillan et al. 1997, Mallet et al. 1998, Jiggins et al. 2001); both processes select against mimetic variants migrating from adjacent areas with distinct warning patterns. In *H. numata,* the evolution of a polymorphic mimicry supergene and disassortative mate preferences could therefore explain the relative lack, compared to other *Heliconius* taxa, of differentiation among polymorphic populations, even across large distances. Furthermore, enhanced gene flow could also cause an increase in effective population size estimates (Slatkin 1987), putatively explaining why polymorphic populations of *H. numata* harbour the highest genetic diversity, and display the highest *Ne* estimates in the entire *melpomene*-silvaniform clade of *Heliconius.*

Alternative processes may of course contribute to the observed patterns. Amazonian and Atlantic populations may differ in other aspects that could also result in differences in genetic diversity. Habitat availability and structure may be different, possibly entailing differences in the maintenance of diversity. The Atlantic Forest is vast in area, but may represent a smaller biome compared to the Amazon, and is isolated from the bulk of the range of *H. numata*, which could result in a population ecology displaying characteristics of peripheral populations with smaller effective population sizes (Eckert et al. 2008). The other *Heliconius* species in the clade have much in common with *H. numata* in terms of habitat and general ecology, yet their niche and life-history specificities and their phylogenetic histories may result in consistent differences with the polymorphic *H. numata* populations. All those specificities may contribute to the observed pattern in which polymorphic Amazonian populations of *H. numata* display high effective population size and a lack of geographic structure in genome-wide genetic variation. Yet this pattern of variation correlates parsimoniously with the evolution of a supergene causing disassortative mating in certain *H. numata* populations, which provides an elegant mechanism explaining their differences with extant and ancestral closely-related lineages. However, we cannot rule out a role for conjectural differences in ecology and geography with all other taxa.

In conclusion, our results show a remarkable contrast in the demography and differentiation of populations within the Amazonian range of *H. numata* compared to closely related taxa and ancestral lineages, as well as with other taxa in the *melpomene*/silvaniform clade. Although those

populations may differ in many uncharacterized ways from all other taxa, one known and consistent difference is the maintenance of inversion polymorphism associated with a specific mating system and selection regime in Amazonian *H. numata*. This distinctiveness of the only widely polymorphic populations in the clade is consistent with the hypothesis that the evolution of a supergene maintained by balancing selection represents a major transition in this lineage, triggering changes in genome-wide patterns of diversity and population ecology over the last 2 million years since its formation. If this hypothesis is correct, the evolution of a locus under balancing selection may therefore feed-back on population ecology and diversification, and consequently on speciation. More work on the determinants of variation in effective population sizes in the *Heliconius* genus is needed to determine the precise impact of the supergene on demography of *H. numata*. We believe that our results emphasize a potential link between genomic architecture, selection and demography, and should inspire future theoretical and modelling studies. Finally, the eco-evolutionary feedbacks between changes in genomic architecture and the ecological parameters of populations are well-known when considering self-incompatibility loci in plants, but may be more common than previously thought. Indeed, our result suggests that balancing selection maintaining structural polymorphisms affecting life-history traits may have a profound influence on species ecology.

Contributions:

MARdC, PJ and MJ designed the study and wrote the manuscript. BH, AVLF, TTT, RRR, KLSB provided the Atlantic samples. CS provided the Colombian samples. MARdC and PJ performed genomic analyses with input from AW. MARdC, PJ, MJ, FPP and MC collected the Peruvian and Ecuatorian samples. MC performed microsatellite analyses and organized fieldworks and butterfly rearing. All authors contributed to editing the manuscript.

Aknowledgements:

This work was funded by grants HYBEVOL (ANR-12-JSV7-0005) and Supergene (ANR-18-CE02-0019-01) from the Agence Nationale de la Recherche and European Research Council Grant MimEvol (StG-243179). We acknowledge the Genotoul and the Montpellier Bioinformatics Biodiversity (MBB) platforms for providing us with calculation time. We thank Dr. Vitor Becker, at the Serra Bonita Reserve (Bahia), Alexandre Soares, at the MN/UFRJ (Rio de Janeiro) and Dr. Marcelo Duarte at the MZ/USP (Sao Paulo) for their contribution to the collection of butterflies in Brazil. Field collections in Colombia were conducted under permit no. 530 issued by the Autoridad Nacional de Licencias Ambientales (ANLA).We are grateful to Marianne Elias, Violaine Llaurens, Quentin Rougemont for comments and discussions. AVLF acknowledges support from Fundação de Amparo à Pesquisa do Estado de São Paulo – (FAPESP) (Biota-Fapesp grants 2011/50225-3, 2013/50297-0) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (421248/2017-3 and 304291/2020-0). KLSB acknowledges the financial support of FAPESP Process # 2012/16266-7. Brazilian specimens are registered under SISGEN (A701768).

Data availability:

The raw sequence data were deposited in NCBI SRA and accession numbers are indicated in Supplementary table 3.

References

1. Alexander DH, Novembre J, Lange K (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**:1655-1664.

2. Beichmann AC, Huerta-Sanchez E, Lohmueller KE (2018). Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms. *Annual Review of Ecology, Evolution, and Systematics* **49**:433–56
3. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**:573-580.
4. Brown KS (1979). Ecologia Geográfica e Evolução nas Florestas Neotropicais. – Univ. Estadual de Campinas, Campinas, Brazil.
5. Brown KS, Benson WW (1974). Adaptive polymorphism associated with multiple müllerian mimicry in *Heliconius numata* (Lepid.: Nymph.). *Biotropica* **6**:205–228
6. Brown KS, Mielke OHH. 1972. The Heliconians of Brazil (Lepidoptera: Nymphalidae). Part II. Introduction and general comments, with a supplementary revision of the tribe. *Zoologica, New York*, **57**:1–40.
7. Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* **10**:195-205.
8. Chouteau M, Arias M, Joron M (2016). Warning signals are under positive frequency-dependent selection in nature. *Proceedings of the National Academy of Sciences of the USA* **113**:2164–2169.
9. Chouteau M, Llaurens V, Piron-Prunier F, Joron M. (2017). Polymorphism at a mimicry supergene maintained by opposing frequency-dependent selection pressures. *Proceedings of the National Academy of Sciences of the USA* **114**: 8325–8329.
10. Davey JW, Chouteau M, Barker SL, Maroja L, Baxter SW, Simpson F, et al. (2016). Major Improvements to the *Heliconius melpomene* Genome Assembly Used to Confirm 10 Chromosome Fusion Events in 6 Million Years of Butterfly Evolution. *G3* **6**:695–708. doi:10.1534/g3.115.023655
11. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**:491–498.
12. Eckert CG, Samis KE, Lougheed SC (2008). Genetic variation across species' geographical ranges: the central–marginal hypothesis and beyond. *Molecular Ecology* **17**:1170–1188.
13. Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey J, Dikow RB, García-Accinelli G, Van Belleghem SM, Patterson N, Neafsey DE, Challis R, Kumar S, Moreira GRP, Salazar C, Chouteau M, Counterman BA, Papa R, Blaxter M, Reed RD, Dasmahapatra KK, Kronforst M, Joron M, Jiggins CD, McMillan WO, Di Palma F, Blumberg AJ, Wakeley J, Jaffe D, Mallet J (2019). Genomic architecture and introgression shape a butterfly radiation. *Science* **366**:594-599.
14. Emsley MG 1965. Speciation in *Heliconius* (Lep., Nymphalidae): morphology and geographic distribution. *Zoologica, New York* **50**:191–254.
15. Faria R, Johannesson K, Butlin RK, Westram AM (2019). Evolving inversions. *Trends in Ecology & Evolution* **34**:239-248.
16. Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, et al. (2012) Genome Sequencing Highlights the Dynamic Early History of Dogs. *PLoS Genetics* **10**:e1004016.

17. Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D (2003) Demography and natural selection have shaped genetic variation in Drosophila melanogaster: a multi-locus approach. *Genetics* **165**:1269-1278.

18. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics* **43**:1031-1034.

19. Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martínez-Aroza J, Oliver JL. (2006) CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics* **7**:446.

20. Hedrick PW, Tuttle EM, Gonser RA (2018) Negative-Assortative Mating in the White-Throated Sparrow. *Journal of Heredity* **109**:223-231.

21. *Heliconius* Genome Consortium (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**: 94–8.

22. Jay P, Whibley A, Frézal L, Rodríguez de Cara MÁ, Nowell RW, Mallet J, Dasmahapatra KK, Joron M. (2018). Supergene evolution triggered by the introgression of a chromosomal inversion. *Current Biology* **28**:1839-1845.

23. Jay P, Chouteau M, Whibley A, Bastide H, Parrinello H, Llaurens V, Joron M. (2021). Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. *Nature Genetics* **53**:288-293.

24. Jiggins C, Naisbit R, Coe R, Mallet J 2001. Reproductive isolation caused by colour pattern mimicry. *Nature* **411**:302–305.

25. Joron M, Wynne IR, Lamas G, Mallet J (1999) Variable selection and the coexistence of multiple mimetic forms of the butterfly *Heliconius numata*. *Evol Ecol* **13**: 721–754.

26. Joron M, Papa R, Beltran M, Chamberlain N, Mavarez J, et al. (2006) A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLoS Biology* **4**:e303

27. 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, ffrench-Constant RH (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**:203–206.

28. Knoppien P (1985) Rare male mating advantage: a review. *Biological Reviews* **60**:81-117.

29. Lenormand T (2002) Gene flow and the limits to natural selection. *Trends in Ecology and Evolution* **17**:183-189.

30. Lenth RV (2016). Least-Squares Means: The R Package lsmeans. *Journal of Statistical Software* **69**:1-33.

31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078-2079.

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

33. Maisonneuve L., Chouteau M, Joron M, Llaurens V (2021). Evolution and genetic architecture of disassortative mating at a locus under heterozygote advantage. *Evolution* **75**:149-165.

480 34. Mallet J, McMillan W, Jiggins C (1998). Estimating the mating behavior of a pair of
481 hybridizing *Heliconius* species in the wild. *Evolution* **52**:503–510.
482 35. Martin SH, Möst M, Palmer WJ, Salazar C, McMillan WO, Jiggins FM, Jiggins CD (2016).
483 Natural Selection and Genetic Diversity in the Butterfly *Heliconius melpomene*. *Genetics*
484 **203**:525-541.
485 36. Mitchell-Olds T, Willis JH, Goldstein DB (2007). Which evolutionary processes influence
486 natural genetic variation for phenotypic traits? *Nature Reviews Genetics* **8**:845–856.
487 37. Muers, M (2009) Separating demography from selection, *Nature Reviews Genetics* **10**:280–
488 281.
489 38. Murray GGR, Soares AER, Novak BJ, Schaefer NK, Cahill JA, Baker AJ, Demboski JR,
490 Doll A, Da Fonseca RR, Fulton TL, Gilbert MTP, Heintzman PD, Letts B, McIntosh G,
491 O'Connell BL, Peck M, Pipes ML, Rice ES, Santos KM, Sohrweide AG, Vohr SH, Corbett-
492 Detig RB, Green RE, Shapiro B (2017). Natural selection shaped the rise and fall of
493 passenger pigeon genomic diversity. *Science* **358**:951-954.
494 39. McMillan W, Jiggins C, Mallet J (1997). What initiates speciation in passion-vine
495 butterflies? *Proceedings of the National Academy of Sciences of the USA* **94**:8628–8633.
496 40. Nadeau NJ, Pardo-Diaz C, Whibley A, Supple M A, Saenko SV, Wallbank RWR *et al*.
497 (2016). The gene *cortex* controls mimicry and crypsis in butterflies and moths. *Nature*,
498 **534**:106–110.
499 41. Nielsen, R., Hubisz, M.J., Hellmann, I., Torgerson, D., Andres, A.M., Albrechtsen, A.,
500 Gutenkunst R, Adams MD, Cargill M, Boyko A, Indap A, Bustamante CD, and Clark AG
501 (2009). Darwinian and demographic forces affecting human protein coding genes. *Genome*
502 *Research* **19**:838–849.
503 42. Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. *PLoS*
504 *Genetics* **2**: e190.
505 43. Rosser N, Phillimore AB, Huertas B, Willmott KR, Mallet J (2012) Testing historical
506 explanations for gradients in species richness in heliconiine butterflies of tropical America.
507 *Biological Journal of the Linnean Society* **105**:479–497.
508 44. Schiffels S, Durbin R (2014) Inferring human population size and separation history from
509 multiple genome sequences. *Nature Genetics* **46**:919-925.
510 45. Saenko SV, Chouteau M, Piron-Prunier F, Blugeon C, Joron M, Llaurens V (2019)
511 Unravelling the genes forming the wing pattern supergene in the polymorphic butterfly
512 *Heliconius numata*. *EvoDevo* **10**:1-12.
513 46. Sheppard PM, Turner JRG, Brown KS, Benson WW, Singer MC (1985) Genetics and the
514 evolution of Muellerian mimicry in *Heliconius* butterflies. *Philosophical Transactions of*
515 *the Royal Society of London, B Biological Sciences* **308**: 433–610
516 47. Slatkin M (1987) Gene flow and the geographic structure of natural populations. *Science*
517 **236**:787-792
518

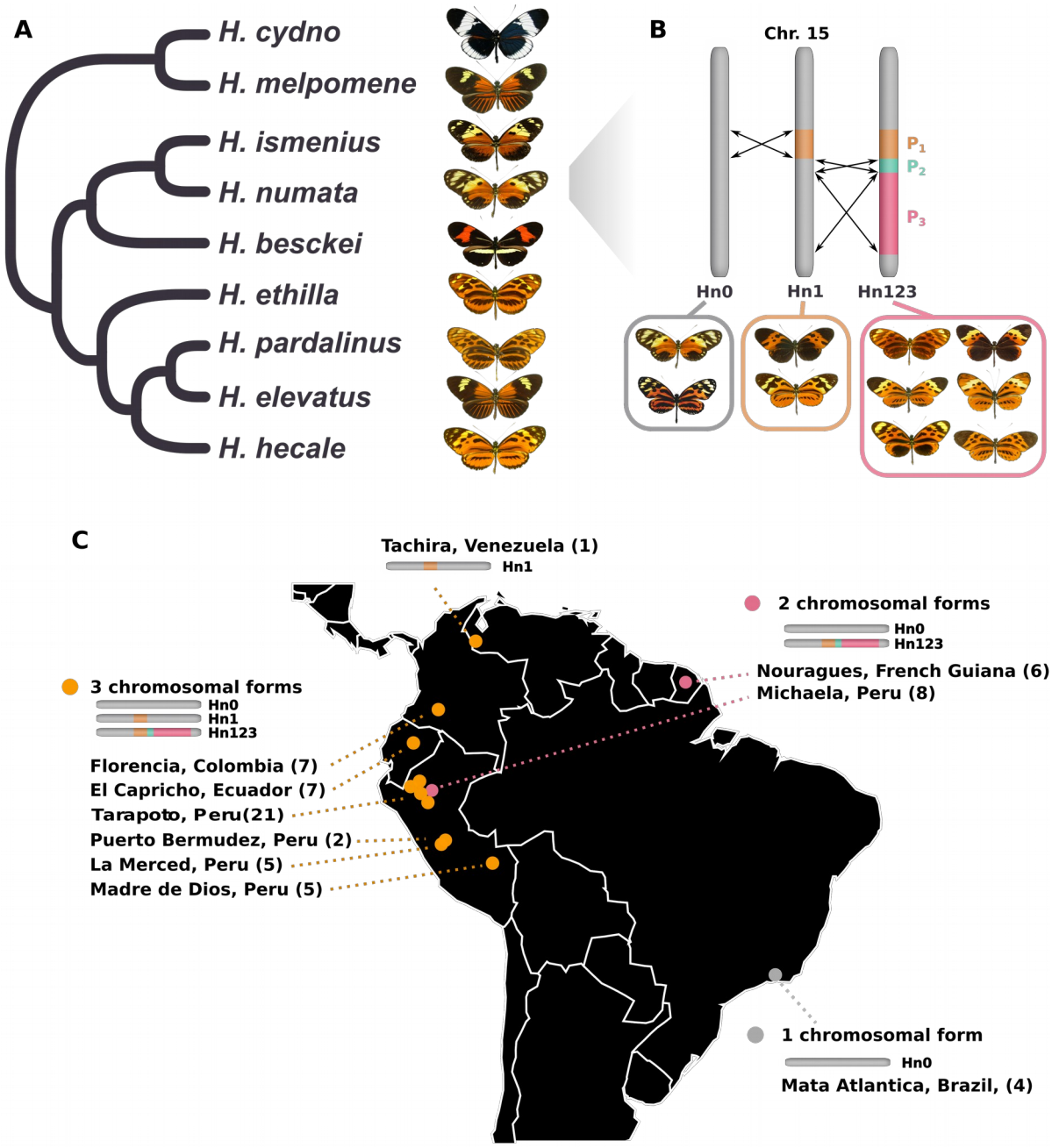

Figure 1 | Genetic and population structure at the P supergene.

A. Schematic phylogeny of the sampled species. It includes all members of the silvaniform clade and two outgroups, *H. melpomene* and *H. cydno*.

B. Schematic description of the genetic structure of the P supergene. Three chromosomal arrangements coexist in *H. numata* and are associated with different morphs.

C. Origin of *H. numata* specimens used for analyses and distribution of chromosome arrangements across the neotropics. Numbers in brackets indicate sampled specimens in each locality (the Tarapoto population lumps several neighbouring subsamples on the map)

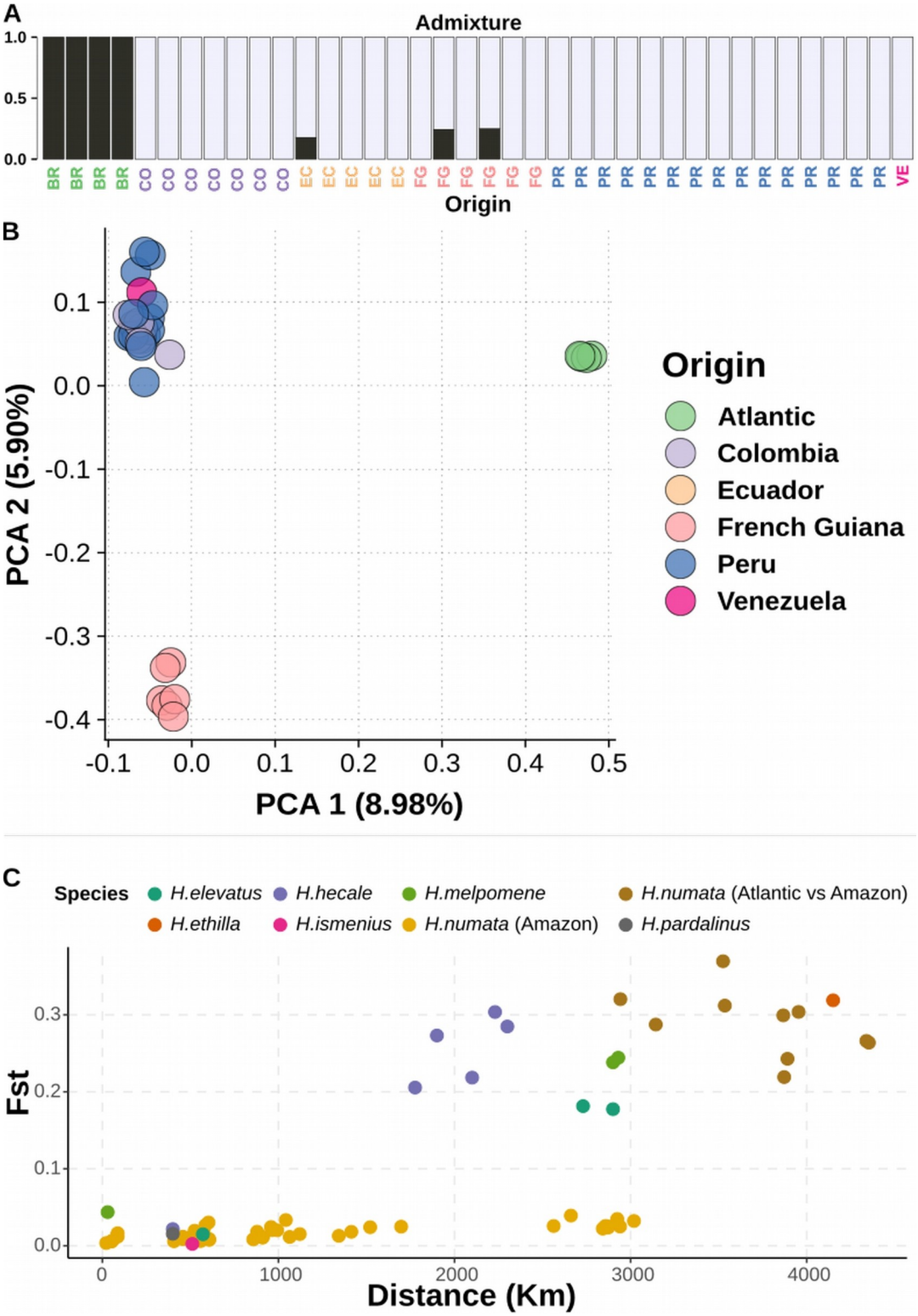

Figure 2 | *H. numata* is characterised by low population structure.
A. Admixture plot for *H. numata*. The optimal cluster number for *H. numata* is two, and it splits *H. numata* into two categories, whereas they come from Atlantic forest or the Amazon. BR=Brazil (Atlantic), PR=Peru, VE=Venezuela, CO=Colombia, EC=Ecuador, FG=French Guiana. **B.** Principal component analysis computed on whole genome SNP. **C.** Relationship between genetic differentiation (Fst) and geographical distance. Fst is measured between morphs/populations of the same species. *H. numata* populations from the Amazon show low isolation by distance when compared to related species.

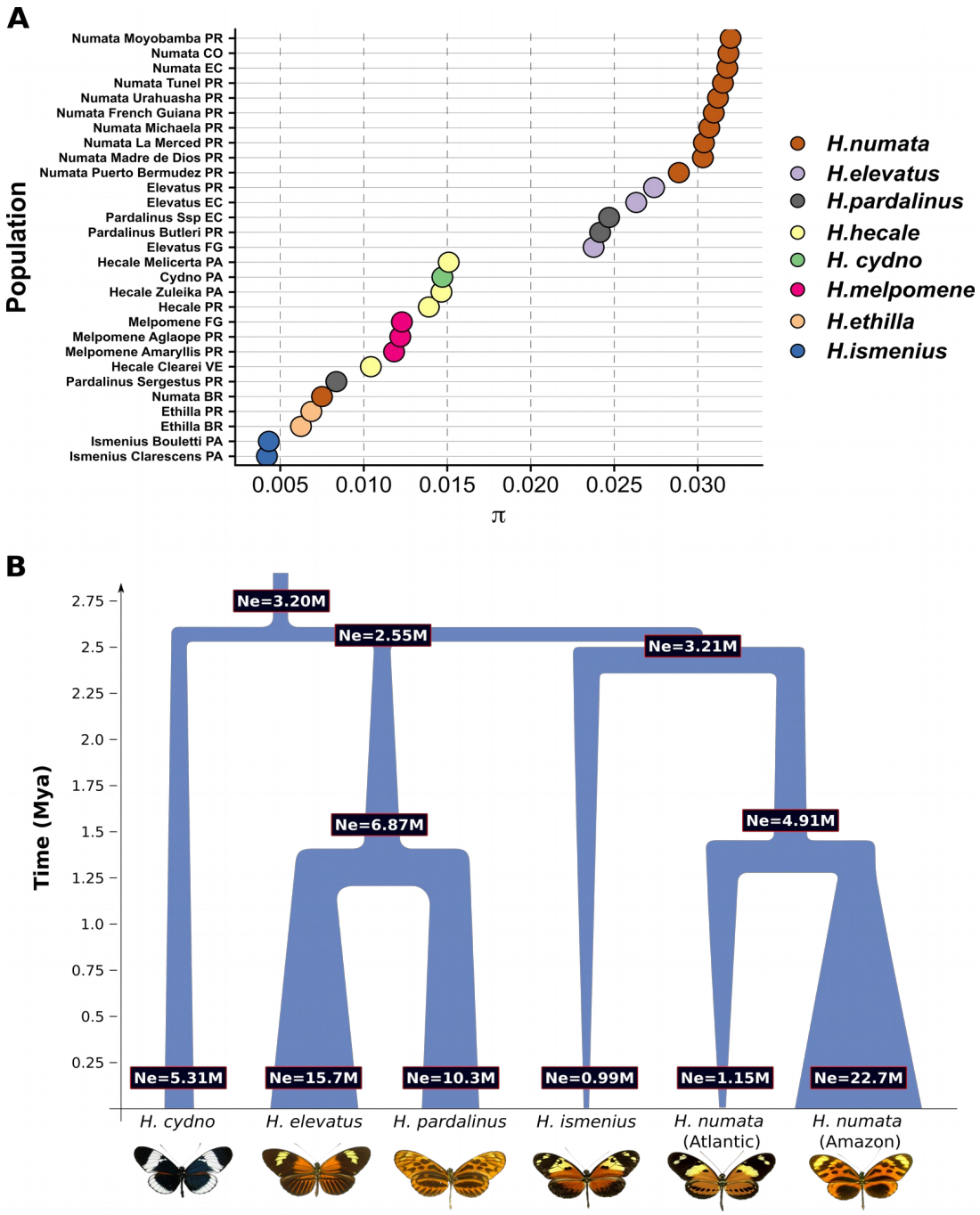

Figure 3 | Variation in present and past effective population size in *Heliconius* species

A. Variation in Pi in several *Heliconius* populations, showing higher genetic diversity in *H. numata* populations from the Amazon than other taxa. Population names indicates their origin as in Figure 2 (e.g. PR=Peru), with the addition of PA=Panama. The *H. numata* population with a lowest diversity is the one from the Atlantic forest (Brazil). **B.** Schematic representation of Gphocs results (presented in table S5-6). Gene flow was modelled but not represented graphically for clarity. showing that Amazonian populations of *H. numata*, which have the P supergene, show a dramatic increase in population size posterior to their split with the Atlantic populations of Brazil, which lack the supergene.

546 **List of Supplementary Materials:**

547 Table S1-6

548 Fig S1-2

549 Text S1

550