Genome-wide admixture is common across the *Heliconius* radiation

Krzysztof M. Kozak[1,2]*, W. Owen McMillan[1], Mathieu Joron[3], Christopher D. Jiggins[1,2]

[1]Smithsonian Tropical Research Institute, Gamboa, Panama.

[2]Department of Zoology, University of Cambridge, Cambridge, United Kingdom of Great Britain and Northern Ireland

[3]Centre d'Ecologie et Evolution, Universite de Montpellier, Montpellier, France

Corresponding author: kozakk {at} si.edu. Twitter: @EvoEcoGen

ABSTRACT

How frequent is gene flow between species? The pattern of evolution is typically portrayed as a phylogenetic tree, implying that speciation is a series of splits between lineages. Yet gene flow between good species is increasingly recognized as an important mechanism in the diversification of radiations, often spreading adaptive traits and leading to a complex pattern of phylogenetic incongruence. This process has thus far been studied in cases involving few species, or geographically restricted to spaces like islands, but not on the scale of a continental radiation. Previous studies have documented gene flow, adaptive introgression and hybrid speciation in a small subsection of the charismatic Neotropical butterflies *Heliconius*. Using genome-wide resequencing of 40 out of 45 species in the genus we demonstrate for the first time that admixture has played a role throughout the evolution of *Heliconius* and the sister genus *Eueides*. Modelling of phylogenetic networks based on 6848 orthologous autosomal genes (Maximum Pseudo-Likelihood Networks) or 5,483,419 high quality SNPs (Ancestral Recombination Graph) uncovers nine new cases of interspecific gene flow at up to half of the genome. However, *f4* statistics of admixture show that the extent of the process has varied between subgenera. Evidence for introgression is found at all five loci controlling the colour and shape of the mimetic wing patterns, including in the putative hybrid species *H. hecalesia*, characterised by an unusual hindwing. Due to hybridization and incomplete coalescence during rapid speciation, individual gene trees show rampant discordance. Although reduced gene flow and faster coalescence are expected at the Z chromosome, we discover high levels of conflict between the 416 sex-linked loci. Despite this discordant pattern, both concatenation and multispecies coalescent approaches yield surprisingly consistent and fully supported genome-wide phylogenies. We conclude that the imposition of the bifurcating tree model without testing for interspecific gene flow may distort our perception of adaptive radiations and thus the ability to study trait evolution in a comparative framework.

INTRODUCTION

Interspecific hybridization and the resulting gene flow across porous species barriers are increasingly recognized as major processes in evolution, detectable at many taxonomic levels (Abbott et al. 2013; Abbott et al. 2016; Schumer et al. 2014; Feliner et al. 2017). This conceptual shift has been facilitated by whole-genome sequencing yielding precise quantitative estimates of introgression (Schrider et al. 2018). However, there are other factors that create the pattern of phylogenetic discordance in genomic data including retention of ancient polymorphism and population structure, incomplete lineage sorting, selection, gene shuffling and duplication. Phylogenomics has massively expanded the potential of comparative molecular biology to elucidate these phenomena in both ancient clades [e.g. birds (Jarvis et al. 2014); mammals (Song et al. 2012)] and recently diverged taxa [e.g. *Virentes* oaks (Eaton et al. 2015), *Drosophila simulans* (Garrigan et al. 2013), *Danaus plexippus* butterflies (Zhang et al. 2014)], but less so at the intermediate level of large, recently-radiated genera. The few genomic studies of recent radiations published to date confirm the prediction of high discordance among independent markers. Analyses of 16 *Anopheles* species and all 48 families of birds (Jarvis et al. 2014) found that no individual gene tree, whether based on protein-coding sequences or large sliding windows, is identical to the species tree topology. This phenomenon appears to be partially attributable to incomplete lineage sorting (ILS) in rapidly radiating groups exposed to novel ecological opportunities, including Darwin's finches speciating after the colonisation of the Galapagos archipelago (Lamichhaney et al. 2015), mammals in response to climatic conditions (Song et al. 2012), and birds after the K/T extinction event (Jarvis et al. 2014; Jarvis 2016).

Discordance due to gene flow between species that are not fully reproductively isolated has been detected in several cases, ranging from humans (Meyer et al. 2012; Sankararaman et al. 2016) to sunflowers (Renaut et al. 2014). In many such taxa genome-wide studies demonstrate both stochastic gene flow across the species barrier, as well as adaptive introgression, where natural selection acts to fix introgressed variants. Alleles at a major locus contributing to the variation in beak shape among Darwin's finches have introgressed between multiple species (Lamichhaney et al. 2015). Similarly,

selection has driven a block of transcription factor binding sites from the ecological generalist *Drosophila simulans* to the specialist *D. secchelia* (Brand et al. 2013), and recent malaria prevention efforts have caused a resistance allele to sweep across *Anopheles gambiae* populations and into *A. colluzzi* (Clarkson et al. 2014; Miles et al. 2017). Nonetheless, the extent and importance of hybridisation in fueling speciation remain hotly contested (Schumer et al. 2014; Feliner et al. 2017; Schumer et al. 2018). Furthermore, gene flow has been less frequently considered than ILS in phylogenetic studies of radiations. One reason is perhaps that modelling gene flow poses greater challenges to computational methods (Wen et al. 2018); for example, almost opposite conclusions were drawn from two methodologically different studies of the same assemblage of *Xiphophorus* swordtail fish (Cui et al. 2013; Kang et al. 2013). The challenge of characterising introgression in adaptive radiations remains open and requires both taxonomic completeness and sophisticated methodological approaches.

The charismatic Neotropical *Heliconius* butterflies have provided some of the most compelling examples of interspecific gene flow occurring pervasively across the genome and at specific adaptive genes (Table 1). The natural propensity of *Heliconius* and the sister genus *Eueides* to produce hybrids in the wild (Mallet et al. 2007; Dasmahapatra et al. 2007) has generated an early interest in the genetic porosity of the species barriers (Beltrán et al. 2002; Bull et al. 2006). Further in-depth studies revealed that ecologically divergent species in the *melpomene-cydno*-silvaniform clade (MCS) can share variation at up to 40% of the genome in sympatry (Kronforst et al. 2006; Nadeau et al. 2012; Nadeau et al. 2013; Mérot et al. 2013; Kronforst et al. 2013; Martin et al. 2013; Martin et al. 2018). Loci responsible for aposematic and mimetic wing patterns are especially likely to be shared between species (Heliconius Genome Consortium 2012; Pardo-Diaz et al. 2012; Enciso et al. 2017; Jay et al. 2018), providing a source of genetic variation in a strongly selected trait *Heliconius heurippa*, also belonging to the MCS clade, remains the best documented case of homoploid hybrid speciation (Salazar et al. 2010; Schumer et al. 2014), in which the adaptive red wing pattern introgressed from *H. melpomene* into the *H. cydno* background contributes to reproductive isolation between species

(Mavarez et al. 2006). Introgression at the *cortex* locus between *H. cydno* and *H. melpomene* facilitated the occupation of atypical mimetic niches in both species (Enciso et al. 2017). More surprisingly, the complex *Dennis/Ray* pattern found in *H. melpomene* and relatives, as well as in *H. elevatus*, is a product of bidirectional adaptive introgression between lineages up to 5 million years apart (Heliconius Genome Consortium 2012; Wallbank et al. 2016). Similarly, introgression of an inversion fixed in *H. pardalinus* triggered the evolution of polymorphic wing patterns in *H. numata* (Jay et al. 2018).

It is virtually unknown whether the phenomena documented in the relatively recently emerged (4.5-3.5MYA) (Kozak et al. 2015) MCS clade are typical of the genus as a whole. Museum collection data suggest that hybridisation happens most frequently among the MCS species (Mallet et al. 2007), which can also be hybridised in captivity (Gilbert 2003). However, many hybrids are also known between *H. erato* and *H. himera*, as well as within the genus *Eueides* (Mallet et al. 2007). Here we generate alignments for 7324 orthologous, protein-coding genes obtained from a comprehensive whole-genome resequencing data set of 40 of the 45 recognized *Heliconius* species, six of the 12 *Euides* species (the most closely related genus) and two outgroup species to investigate the prevalence of hybridisation in *Heliconius*, quantify its extent, and compare the processes producing discordance at several loci. We demonstrate varied amounts of phylogenetic incongruence across the genus, related to heterogenous levels of interspecific gene flow. Instances of hybridisation at adaptive loci and more broadly across the genome, are found across the radiation, although they are far more frequent in the relatives of *H. melpomene*.

Our study dissects the evolution of two genera across a continent to demonstrate that gene flow can be a ubiquitous and significant force in shaping an adaptive radiation, and to provide first estimates of the probability of such gene flow and adaptive introgression. We show that a misleadingly well-supported and resolved tree can be recovered despite incongruence, and that previously unknown, complex hybridisation event can thus be missed. We propose that phylogenetic networks should become *de rigour* tools in the study of adaptive radiations.

RESULTS AND DISCUSSION

A genomic tree for the heliconiines

We first constructed a bifurcating phylogeny for our genome-wide data. Mapping genome-wide short read data from 144 individuals in 45 species (S1 Table) to the *H. melpomene* reference did not result in full coverage of the distantly related species (S2 Table), but we were able to recover 7264 high quality, orthologous CDS alignments with less than 4% missing data (S3 Table). Filtering for autosomal exome sites with biallelic, non-singleton SNPs without missing data produced a 122,913 bp supermatrix. This matrix generated a Maximum Likelihood tree that is resolved with full bootstrap support (S1 Figure), except for uncertain placement of the *H. telesiphe/hortense/clysonymus* clade (bootstrap support 62/100). This concatenation tree is not substantially different from the results obtained by either ASTRAL or MP-EST multispecies coalescent methods (Fig. 1), which model the species tree from the distribution of gene trees. Both MSC methods also yield well resolved phylogenies and, in the case of ASTRAL, completely supported topologies (Fig. 1). The MSC trees differ from each other at only two relatively recent splits, while the concatenation phylogeny differs from ASTRAL at three out of 56 (Fig. 1).

At one level, these well resolved trees clarified some uncertainties from previous work (Beltran et al. 2007; Kozak et al. 2015), including placement of the *H. hecuba* and *H. egeria* groups, relations in the *H. sapho* clade, and the position of *H. besckei*. Nonetheless, similar to other large phylogenetic studies (Brawand et al. 2014; Jarvis et al. 2014; Fontaine et al. 2014), none of the individual gene trees showed exactly the same topology as the species tree. making us realise that the well-supported bifurcating trees do not fully represent the underlying signal in the genomes within this clade.

Genome-wide phylogenetic incongruence

Rapid radiations of closely related species are likely to show considerable phylogenetic incongruence due to incomplete lineage sorting and introgressive hybridization. This is especially likely to be the

case in *Heliconius*, where population sizes are large (Flanagan et al. 2010), speciation events are clustered in time (Kozak et al. 2015), inter-specific hybrids are found in the wild (Mallet et al. 2007), and there is strong evidence for extensive introgression of closely related sympatric species (Martin et al. 2013; Martin et al. 2018). Indeed, despite the apparently well-supported phylogeny, we found considerable phylogenetic incongruence across the genome in our data. For example, Robinson-Foulds pairwise distance was 0.745 out of 1.0 for the 6848 autosomal phylogenies and 0.699 for the 416 loci on the sex-linked Z chromosome, indicating that any two gene trees are likely to contain multiple differing partitions. Among the 56 nodes between species and major subspecies, less than a half (26) are resolved in an autosomal Majority Rule Consensus tree (S2-S4 Figures). The relative tree certainty (TCA) of the consensus is 0.322 on a 0-1 scale, whereas many supported branches score low on the Information Criteria (IC/ICA), indicating that there is one (IC) or multiple (ICA) alternative groupings found at high frequency among the gene trees (Salichos & Rokas 2013) (S2 Figure). Brower and Garzon-Orduña (2017) suggested that the phylogenetic incongruence among Heliconiini is an artifact of missing data. This is unlikely, as the 7324 nucleotide matrices recovered here are nearly complete (>96%) and only 0.87% of the sites were removed as incomplete or ambiguous (S3 Table). Phylogenetic studies using modern statistical methods are typically robust to far higher levels of missing data (Wiens & Morrill 2011; Roure et al. 2013).

It has been suggested that sex-linked markers might be more reliable than autosomes for phylogenetic reconstruction, due to faster coalescence at lower effective population size (¾ of the autosomal N_e at the Z), and reduced likelihood of hybridisation as a result of Haldane's Rule (Zhang et al. 2013; Fontaine et al. 2014). We found that there is indeed significantly greater concordance between species tree and the Z chromosome trees, as compared to autosomal trees (gene tree-species tree triplet distance: $10.3*10^6$ *vs* $12.5*10^6$, Wilcoxon's test $p=3*10^{-11}$) (S3 Figure). Notably, many nodes within the MCS clade are resolved among Z chromosome trees, and the *melpomene/cydno* clade is monophyletic. This agrees with genomic studies of house mouse subspecies (White et al. 2009), fruit flies (Garrigan et al. 2012; Pease & Hahn 2013) and *Anopheles* mosquitoes (Lee et al. 2013; Fontaine

et al. 2014) showing deeper coalescence and lower levels of phylogenetic discordance at the X chromosomes. A comparison of gene tree discordance levels, and coalescent trees between autosomes and the Z demonstrates that the sex chromosome in *Heliconius* is likely subject to slightly lower levels of gene flow (S4 Figure). Nonetheless, the Z-linked consensus did not resolve the relative positions of the major clades unequivocally (S2-S3 Figures) and there were roughly similar levels of conflict with the species tree between sex-linked and autosomal loci (S5 Figure). Thus, even the Z chromosome yields inconsistent the phylogenic estimates (S4 Figure). This may be a result of an ancient introgression of the entire chromsome, since with our larger species sampling we confirm previous demonstration of Z chromosome incongruity between the MCS clade and the Silvaniform group containing *H. ethilla, H. hecale* and *H. pardalinus* (S6 Figure) (Zhang et al. 2016).

Similarly, the whole mitochondrial phylogeny was strongly conflicted with the multispecies coalescent trees (Fig. 1), demonstrating that organellar history is unlikely to be a good proxy for phylogeny at the species level (Hurst & Jiggins 2005). Generally, the same branches tended to be short and unsupported in concatenation and coalescent analyses of mitochondrion-, Z- and autosome-linked markers (Fig. 1, S1-S3 Figure, S6 Figure). This corroborates the idea that some of the diversification events among Heliconiini occurred nearly simultaneously and are unlikely to be resolved in a bifurcating tree.

Hybridisation between species has been common throughout the radiation of *Heliconius*

We next sought to document the extent and degree of discordance across the phylogeny, and distinguish between incomplete lineage sorting and gene flow. We did this in two ways, first modeling the extent of hybridization in the history of the radiation in the software PhyloNet, using the Maximum Pseudo-Likelihood network algorithm for distinguishing admixture from Incomplete Lineage Sorting based on gene tree topologies and branch lengths (Yu & Nakhleh 2015). Here, an inference of speciation history networks from 6848 gene trees revealed a pattern of gene sharing within all major clades of the radiation (Fig. 2). This was particularly true within the *H. melpomene,*

H. cydno and Silvaniform clade (MCS), where our approach supports previous work documenting extensive admixture, including the hybrid speciation of *Heliconius heurippa* (Mavarez et al. 2006; Salazar et al. 2008; Salazar et a. 2010), admixture between *H. cydno/timareta* and subspecies of *H. melpomene* (Nadeau et al. 2013; Martin et al. 2013; Enciso et al. 2017), and the exchange between *H. melpomene* and *H. elevatus* (Heliconius Genome Consortium 2012; Wallbank et al. 2016). This result made us confident that outcomes of this analysis can be trusted.

In the small clades containing *H. egeria*, *H. hecuba* and *H. aoede*, we found similar evidence for multiple instances of admixture, particularly between all species in the *H. hecuba* quartet (Fig. 2B). Gene flow is also apparent between the ancestors of modern clades, including a potential hybrid origin of *H. aoede*. This finding explain the unusual combinations of morphological and ecological traits in these species, which used to be classified in separate genera (Brown et al. 1981; Brower 1994) prior to molecular evidence (Beltran et al. 2007; Kozak et al. 2015). Finally, in the large group of *H. erato* (*sensu* Brown 1981)we detected admixtures between the triplet of (*H. telesiphe*,(*clysonymus*, *hortense*)) and both *H. erato* and *H. sara* clades (Fig. 2A). Notably, there are fewer instances of reticulation than in all the other groups consistent with other evidence (Fig. 4, S9 Figure). The most likely network estimated for the genus *Eueides* contains an unusual reticulation pattern leading to none of the included taxa (S1 Figure). We suspected that this pattern may be a sampling artefact, as only half of the species are included with one individual each, and divergence from the reference genome makes the *Eueides* alignments less complete and reliable.

Although MPL networks are currently the most robust and computationally efficient tools to model gene flow in face of ILS (Wen et al. 2018), it is unknown if the estimates are sensitive to errors in gene tree estimation and missing data as in case of *Eueides*. To account for this possibility we also used the Ancestral Recombination Graph algorithm implemented in TreeMix (Pickrell & Pritchard 2012), which analyses genome-wide variation under a Gaussian drift model, interpreting deviations from expected differentiation as evidence for gene flow. When applied to the SNP supermatrix, TreeMix again detects known events in the MCS clade, confirming the robustness of the technique,

even when applied at the relatively deep level of a genus (Fig 3).

The ARG shows gene flow from *H. hecalesia* into the *telesiphe* clade, and from the *erato* lineage into *H. hecalesia* (Fig. 3). This suggests a hybrid origin for *Heliconius hecalesia* (Fig. 2A). In contrast, and somewhat surprisingly, we find no evidence for a hybrid origin of *H. hermathena*. This species, found in the white sand forests of Brazil, has a wing pattern resembling a combination of the zebra-patterned of *H. charithonia* and the red-banded *H. erato hydara* (Brown & Benson 1977; Mavárez & Linares 2008; Mavárez et al. 2006).

In the small clades including *H. egeria*, *H. hecuba* and *H. aoede*, we find evidence for pervasive admixture, particularly between all species in the *hecuba* quartet (Fig. 2B). In contrast, TreeMix detects only a single gene flow event from *H. aoede* into the *egeria/hecuba* clade (Fig. 3). Gene flow is also apparent between the ancestors of modern subgenera, including a potential hybrid origin of *H. aoede*, which may explain the unusual trait combinations in these species that were sometimes classified in separate genera (Brown et al. 1981; Beltran et al. 2007; Kozak et al. 2015) (Fig. 2, 3).

The Treemix ARG and PhyloNet network approaches do not uncover an exactly identical set of introgression results, as the ARG does not detect the exchanges in and between the small *H. egeria* and *H. hecuba* clades, or some of the events previously documented between species of the Silvaniform group (Jay et al. 2018) and *H. melpomene* (Zhang et al. 2016). The discrepancies are expected between two widely different techniques, and the Treemix algorithm assumes that the underlying sequence of events was largely tree-like (Pickrell and Pritchard 2012). Although Treemix is designed for populations and less commonly used to examine divergent species (but see Zhan et al. 2014), it infers a basic tree similar to that found by the phylogenetic approaches that do not model introgression (Fig. 1).

Distributions of allele frequencies show that the extent of introgression between species varies between clades within *Heliconius*. When all possible combinations of taxa are examined, most of the quartets in the MCS clade (22 species, 3309790 SNPs) show some admixture ($f4<0$; Z-score<-4.47 after Bonferroni correction for the number of lineages) (Fig 4). In contrast, most *f4* values computed

from the *H. erato* group (20 species, 3152880 SNPs) are not significant and thus indicate a relatively better fit to a simple bifurcating tree (Peter 2016). This summary statistic allows a crude comparison of levels of admixture across the genus, but it remains computationally difficult to examine gene flow between more than a handful of taxa using these approaches (Martin et al. 2015; Pease & Hahn 2015).

Adaptive introgression at the wing pattern loci

A characteristic aspect of *Heliconius* evolution is the adaptive introgression of five colour pattern loci, corresponding to a small number of protein-coding genes (Table 2) (reviewed in Kronforst and Papa 2015). For many of these genes more recent work has identified intervals that are associated with specific aspects of color and pattern that are thought to contain regulatory regions responsible for differential expression (Table 2). Introgression driven by selection for Mullerian mimicry has been documented extensively (Salazar et al. 2010; Pardo-Diaz et al. 2012; Heliconius Genome Consortium 2012; Wallbank et al. 2016; Zhang et al. 2016; Jay et al. 2017), but only in the *H. melpomene* and Silvaniform (MCS) clade of *Heliconius*, which includes less than a third of all species. By comparing the topologies at the colour pattern loci for nearly all species in the genus against the overall species tree, we provide a uniform framework in which to gauge the amount of introgression across around these functionally important regions. Topologies around color pattern loci differed from the species tree (p<0.001, SH test) and in many cases, primarily at the *optix* and *cortex* loci, showed multiple departures from the expected topology (Table 1). The majority of the differences are found in the MCS clade, where introgression reaches considerable complexity across genomic and geographic regions (Wallbank et al. 2016; Enciso et al. 2017). Again, we verify our approach by recovering the known signals the *optix* and *cortex* loci (Table 1), but it is the first time that introgression is demonstrated in other genomic intervals.

In contrast, the *H. hecalesia/clysonymus/hortense* is the only case of wing pattern introgression found among the 19 species of the *H. sara/erato/telesiphe* subgenus. No pattern introgression is found among the three smaller clades (for example Fig. 6), despite evidence for gene flow in other regions

(Fig. 2, 3). Although we cannot rule out an influence of sampling bias, as substantially more individuals have been sequenced in the MCS group, our data nonetheless indicate that introgression around color pattern loci is more pervasive in this clade.

We find new evidence for gene flow at the *cortex* locus (Nadeau et al. 2016) which is responsible for the diverse white and yellow patterns across the genus (Sheppard et al. 1985; Kronforst & Papa 2015; Nadeau et al. 2014). At a region previously highlighted (Nadeau et al. 2012) (310-330kbp in scaffold HE667780) *H. melpomene* and *H. timareta* alleles cluster with Silvaniforms (aLRT > 0.95); further downstream *H. melpomene, H. timareta* (570-620 kbp) and *H. cydno* (590-620 kbp) share alleles, as well as a cluster of sequences similar between *H. elevatus/pardalinus* and *H. melpomene/cydno* (570-600 kbp). This last association is also observed in the 670-690 kbp window containing *cortex* (Nadeau et al. 2014).

At the *WntA* locus, we again find topologies indicating Amazonian *H. melpomene* introgression into *H. timareta* and *H. elevatus*. In the *WntA* (Martin et al. 2012) windows (450-490 kbp), sequences of *H. heurippa* cluster with *H. cydno*, upholding the view that the former arose by hybrid speciation from a yellow-patterned race of *H. cydno/timareta*.

Most of the variation in the *optix* region is consistent with the genome-wide lack of resolution in the *H. melpomene/cydno*/Silvaniform clade and confirms known events. The greatest number of discordant branches are among the *H. melpomene/cydno* clade at 360 to 380 kbp (Fig. 6), the section controlling both *H. melpomene* (Wallbank et al. 2016) and *H. erato* red ray patterns (Supple et al. 2013; van Belleghem et al. 2017). Intriguingly, alleles from *H. hecale clearei* cluster with the *H. pardalinus/H. elevatus* sequences in eight windows (Fig. 6), perhaps related to the complete loss of orange patterning in this uniquely black and white Silvaniform.

Heliconius stand out as having multiple examples of introgression of unlinked loci enabling the rapid evolutionary development of complex patterns, which comprise a patchwork of elements sometimes derived from different sources (Table 1)(Wallbank et al. 2016). A large number of genomic studies of interspecific gene flow have found introgressions of small genome regions driven by natural selection

for critical adaptations, such as the hypoxia resistance *EPAS1* haplotype (Denisovans → anatomically modern Tibetans) (Huerta-Sánchez et al. 2014), the *Vgsc-1014F* insecticide-resistance mutation (*Anopheles gambiae* → *A. colluzzi*) (Clarkson et al. 2014; Norris et al. 2015), or the *ALX1* alleles determining diverse beak shapes among Darwin's finches (*Geospiza*) (Lamichhaney et al. 2015). Whereas adaptive introgression has been localised to a single part of the genome in all but one other animals studied to date (but see Stryjewski & Sorenson 2017), it is clear that *Heliconius* wing patterns involve introgressions at multiple loci and between different combinations of species. For instance, *H. elevatus* derives *WntA* and *optix* alleles (Table 1) (Wallbank et al. 2016) from Amazonian *H. melpomene*, but some of the *WntA* haplotype appears to be transmitted from *H. cydno*, possibly *via H. melpomene*. Similarly, fragments of the *optix, WntA* and *cortex* sequences in *H. elevatus* are derived from either *H. melpomene* or *H. cydno*.

Mosaic genome in *Heliconius hecalesia*

We identify the first cases of gene flow among distantly related species in the *H. erato* clade, demonstrating that admixture has not been limited to the well-studied MCS group. We infer two instances of gene flow. into both the ancestor of *H. hortense* and *H. clysonymus,* and into *H. hecalesia*. The position of the (*H. telesiphe*,(*hortense*,*clysonymus*)) triplet (THC) is the only unsupported branch in the genome-wide phylogeny (Fig. 1; S2 Figure), and no specific placement is found in a majority of gene trees (IC=0; S3 Figure). This uncertainty is not a result of coalescent stochasticity, as both MPL networks and TreeMix account for ILS and still recover signals of admixture from both the clades of *H. erato* and *H. sara* into THC (Fig. 2A, 3). *H. hecalesia* latter has especially complex ancestry, and although usually grouped with *H. erato* in simple trees (Fig. 1) (Kozak et al. 2015), appears to be nearly equally diverged from *H. erato*, *H. telesiphe* and *H. sara* (Fig. 5). Results of network modelling are corroborated by the *D* statistic of admixture (Durand et al. 2011), which is highly positive and statistically significant for all tests where *H. hecalesia* is the recipient of admixture from the *H. clysonymus* and *H. sara* clades (Table 3). However, consistent with the pattern of genome-wide

variation, there is evidence for stronger gene flow between similarly patterned *H. hecalesia* and *H. clysonymus* (D=0.35; p<0.0001) or *H. hortense* (D=0.38; p<0.0001) than the dissimilar *H. sara* (D=0.17). At the higher phylogenetic level there is also evidence for gene flow between the THC clade and *H. hecalesia*, but not *H. erato* (Table 3), raising the possibility that admixture took place between the ancestors of modern clades, leading to incongruence deep in the tree (S1 Figure).

For the first time we find evidence outside of the MSC clade for introgression of multiple colour pattern alleles, from *H. hecalesia* into the sympatric THC lineage, possibly due to selective pressure on mimetic appearance. *H. clysonymus* and *H. hortense* display the unusually wide red/orange bar on the hindwing, absent from their relative *H. telesiphe*, but shared with the distantly related *H. hecalesia* (Fig. 5). Based on the phenotype we predict that, at the key wing patterning loci (Table 2), sequences from these three species should cluster together to the exclusion of *H. telesiphe*. Sequences from the three species indeed cluster exactly at the 360-380kbp interval of scaffold HE670865 (aLRT>0.95; p<0.001, SH test), which aligns to the specific region of the *D* locus controlling the red pattern in *H. erato* (Supple et al. 2013) (Fig. 6). This indicates that the alleles governing the red pattern in the three species are more similar than expected from the genome-wide phylogenies. As predicted if the sequences were introgressed after species divergence, the alleles from the three phenotypically similar butterflies appear younger (2.3-2.1 MA) in RelTime estimates than those of *H. telesiphe* (2.9MA), and much more recent than the original split between *H. hecalesia* and the THC clade (6.28-4.18 MA) (Kozak et al. 2015).

Whereas *optix* colours the band red (Zhang et al. 2017), the shape of bands is typically regulated by *WntA* (Mazo-Vargas et al. 2017), as well as by the *Ro* locus (Nadeau et al. 2014). As predicted from the shared phenotype of a wide hindwing band, we find that *H. hecalesia/clysonymus/hortense* cluster both at the *WntA* interval (HE667780: 450-490 kbp) (Martin et al. 2012), and in the section of *Ro* containing eight SNPs associated with pattern shape in *H. erato* (HE671554: 10-30kbp). Consistently, sequences of the phenotypically different *H. telesiphe* are divergent.

Adaptive gene flow between the three species is plausible, as *H. clysonymus* x *H. hecalesia* and *H. hortense* x *H. hecalesia* hybrids are known from the wild (Mallet et al. 2007), and *H. hecalesia* is sympatric with the other two species in parts of its range (Rosser et al. 2012). The subtle changes in the wing morphology of *H. hecalesia* across its range appear driven primarily by sexually-dimorphic mimicry of species in the Ithomiinae, with a wider hindwing band found further away from the Equator (Brown & Benson 1975). It seems likely that the precise mimicry between the distantly related *H. hecalesia* and *H. clysonymus/hortense* has resulted from localised introgression of regulatory loci in sympatry, similar to that seen between *H. melpomene/elevatus* (Wallbank et al. 2016).

Neither concatenation nor coalescent trees adequately represent species history

There has been a marked shift over recent years away from phylogenetic methods that involve concatenation of data towards approaches that involve coalescent modelling. The proposed methods for inferring a species tree under the coalescent by modelling the incomplete sorting of loci represent a great improvement on the assumption that there is a common evolutionary history across all genome regions (Liu et al. 2010; Mirarab et al. 2014; Heled & Drummond 2010). Across the tree of life, from birds (Jarvis et al. 2014; Reddy et al. 2017) and mammals (Song et al. 2012) to fungi (Shen et al. 2016), treatment of individual gene trees under multispecies coalescent methods has yielded substantially different results to simple concatenation. However, both approaches still impose a bifurcating tree on the data and imply that speciation consists of a series of splits (Hahn & Nakhleh 2016). When applied to the adaptive radiation of *Heliconius*, both philosophies appear to produce results that fail to fully describe the evolutionary history of the species. Network modelling clearly demonstrates that introgression has happened throughout the evolution of the genus, and yet this process could easily be overlooked even with many state of the art phylogenetic methods. Appropriate reconstruction of speciation history needs to incorporate approaches that test for admixture more routinely (Long & Kubatko 2017; Pease & Hahn 2015). The main appeal of studying adaptive

radiations is their power for analyzing trait evolution in a comparative framework, and a growing number of studies are looking at several *Heliconius* characters through this lens (Mallet et al. 2007; Briscoe et al. 2013; Rosser et al. 2015). Commonly, trait histories are best represented by exactly those regions that show high levels of introgression, and a comparative approach that used a single bifurcating species tree would give highly incomplete results. To accurately reflect the hidden uncertainties in the phylogeny of this exciting group, and to capture the potential of traits to be shared between species by genome-wide introgression, future work needs to utilize novel approaches that reflect the non-bifurcating reality of evolving genomes (Hahn & Nakhleh 2016; Bastide et al. 2017).

METHODS

Samples and DNA sequencing

Whole genomes of 11 species were re-sequenced for the first time: *Heliconius atthis*, *H. antiochus, H. egeria, H. leucadia, H. peruvianus, Eueides aliphera, E. lampeto, E. lineata, E. isabella, E. vibilia, Agraulis vanillae. H. antiochus* and *H. ricini* were collected in Bakhuis Mountains, Suriname. Other data were obtained from previous studies (Heliconius Genome Consortium 2012; Kronforst et al. 2013; Martin et al. 2013; Supple et al. 2013; Briscoe et al. 2013; Martin et al. 2016; Nadeau et al. 2016; Wallbank et al. 2016; Enciso et al. 2017; Jay et al. 2018). To enhance coalescent modelling by representing genetic diversity (Edwards et al. 2016), samples were chosen from distant populations and diverse wing pattern races as available. The total dataset included 145 individuals across 40/45 species of *Heliconius*, 6/12 *Eueides* and the monotypic *Dryadula* and *Agraulis* (S1 Table). Sequencing was performed using the Illumina technology (S1 Table) with 100 bp paired-end reads, insert sizes of 250 - 500 bp and read coverage from 12x to 110x. In case of 13 new samples, DNA was extracted with the DNeasy Blood and Tissue kit (Qiagen) from 30-50 µg of thorax tissue homogenised in buffer ATL using the TissueLyser (Qiagen); purified by digesting with RnaseA (Qiagen); and quantified on a Qubit v.1 spectrophotometer (LifeTechnologies). The Beijing Genomics Institute constructed and sequenced whole-genome libraries on a HiSeq 2500 with 500 bp insert size to ~50x coverage (S1 Table).

Read mapping and genotyping

Raw reads were checked using FastQC v0.11 (Andrews 2014) and mapped to the *H. melpomene melpomene* reference (Heliconius Genome Consortium 2012) with BWA v6(Li & Durbin 2009), followed by refinement with Stampy v1.0.18 (Lunter & Goodson 2011). Aligner parameters were based on empirical tests (Nadeau et al. 2013; Davey 2013) and the age of divergence from the reference (Kozak et al. 2015) (S2 Table). Alignments were sorted with Samtools(Li et al. 2009), de-duplicated with Picard v1.112 (Fennell 2010) and re-aligned in Genome Analysis Toolkit v3.1

(GATK) (McKenna et al. 2010; DePristo et al. 2011). SNPs were called separately across samples at sites with coverage >4x and quality >20 using the GATK UnifiedGenotyper(van der Auwera et al. 2013). The liberal values that increase calling sensitivity were chosen as appropriate for a phylogenetic analysis, where a few individual SNPs are not viewed as evidence in isolation. Species genotypes were merged using Bcftools v1 (Li et al. 2009) and assessed with an in-house Python script (Martin et al. 2013) (evaluateVCF-03.py (Martin 2017)). We identified 126,865,683 individual SNPs, including 5,483,419 in the exome. The autosomal matrix of exonic, biallelic, non-singleton SNPs genotyped in all individuals contained 122,913 variants (Supplementary Methods).

Commands for genotyping and phylogenetic software are listed in Supplementary Methods.

Exome alignments and gene trees

Protein-coding genes can be effectively treated as discrete markers for multilocus phylogenetics(Edwards et al. 2016) with lower level of homoplasy than the non-coding sequence(Brawand et al. 2014). We minimized paralogy by narrowing the gene set to 1:1:1 orthologs between *H. melpomene*, *Danaus plexippus* (Zhan et al. 2011) and *Bombyx mori* (Xia et al. 2004) identified by OrthoMCL (Li et al. 2003; Heliconius Genome Consortium 2012). 6848 autosomal gene alignments, excluding those found on the colour pattern scaffolds (Heliconius Genome Consortium 2012), and 416 Z-linked single-copy genes were extracted using the tabular "calls file" format [gene_fasta_from_reseq.py (Martin 2017)]. TrimAl v1.2 (Capella-Gutiérrez et al. 2009) was used to remove the taxa for which 50% of residues did not overlap with at least 50% of the other sequences, while uninformative fast-evolving regions were excluded by Block Mapping and Gathering with Entropy (BMGE) (Criscuolo & Gribaldo 2010). Individual ML gene trees were estimated in FastTree v2.1 (Price et al. 2010) with parametric aLRT nodal support (Anisimova & Gascuel 2006).

Incongruence in the data

To assess the level of topological variation in the gene trees, the average normalised Robinson-Foulds

distance (RF) (Robinson & Foulds 1981) between all pairs of phylogenies was computed in PAUP* v4 (Swofford 2002). The calculation was repeated after trimming to 57 individuals representing species or highly distinct subspecies (e.g. the three geographic clades of *H. melpomene*), thus eliminating the noise from lack of intraspecific resolution. The 57 samples were chosen based on the quality of the genotype calls (S1 Table). In addition, the average distance between all possible triples of taxa represented in full gene trees - a measure similar to the RF - was calculated in MP-EST (Liu et al. 2010). To identify incongruent nodes, 50% Majority Rule consensi were calculated from the trimmed gene trees and the relative support for branches was evaluated under the Internode Certainty criteria (IC/ICA/TCA), which compare the support for a branch and for all alternatives found in the distribution (Salichos & Rokas 2013; Salichos et al. 2014). The effect of poor resolution in some gene trees was accounted for by repeating the procedure with the 1000 most resolved and supported phylogenies. To quantify the gene-species tree disagreements, we computed what proportion of gene trees contain the quartets found in the ASTRAL-III phylogeny (*–t1*) (Sayyari & Mirarab 2016).

Species trees

Naïve "total evidence" phylogenies were estimated from concatenated genome-wide SNPs in RAxML v8 with 100 bootstrap replicates, GTR+Γ model and ascertainment bias correction (Stamatakis 2014). The history of the matriline was approximated from the whole-mitochondrial alignment with partitions determined by PartitionFinder v1.1 (Lanfear et al. 2012). To verify a previous dating of the radiation(Kozak et al. 2015), an ultrametric Minimum Evolution tree was estimated from the autosomal SNPs using relative rate comparison in RelTime (Tamura et al. 2012; Tamura et al. 2013), constraining intervals around the age of the *Heliconius-Eueides* (17.0 - 20.0 MA) and *Heliconius-Agraulis* (25.0 - 28.0 MA) as determined for the Nymphalidae phylogeny (Wahlberg et al. 2009).

Two Multispecies Coalescent (MSC) approaches were used to estimate the species tree under the assumption of Incomplete Lineage Sorting. MP-EST v1.4 maximises a pseudo-likelihood function

over the distribution of taxon triples extracted from gene tree topologies(Liu et al. 2010). 145 individual tips were assigned to species *a priori* (S1 Table), treating Western and Eastern clades of *H. melpomene* and *H. erato* as distinct (Nadeau et al. 2013; van Belleghem et al. 2017). Bootstrap support was evaluated by repeating 100 times with random samples of 500 gene trees. Since MP-EST may be misled by errors in gene tree reconstruction(Mirarab & Warnow 2015), we compared the results with ASTRAL-III, a fast quartet method that deals with polytomies and low support values in the input (Zhang et al. 2017).

Modelling hybridisation

Following the identification of problematic nodes based on ICA and MSC trees, we applied three radically distinct approaches to disentangle ILS from hybridization (Hahn & Nakhleh 2016; Peter 2016; Scornavacca & Galtier 2016). First, we determined the Ancestral Recombination Graph (ARG) in TreeMix v1.13 by identifying the pairs of taxa sharing more than the expected proportion of allelic variation (Pickrell & Pritchard 2012). Relations between major lineages and species were inferred from allele frequency data computed in PLINK! v2 (Purcell 2009; Chang et al. 2015) after Linkage Disequilibrium prunning (Gaunt et al. 2007), removing sites with less than 95% complete data, and grouping SNPs into blocks of 100. Models were fitted with zero, 10 or 20 migration events, treating *H. aoede* as an arbitrary outgroup for rooting.

Second, we tested the fit of the data to the bifurcating tree ideal by calculating the *f4* statistic (Reich et al. 2009), a difference in allele frequencies of two taxa pairs that can be interpreted as a measure of treeness (Peter 2016), similar to the *D* statistic (Durand et al. 2011). The *f4* was calculated for all possible quartets of taxa in each clade using the *fourpop* routine in TreeMix on the SNP data described above, determining the statistical significance of the results by jacknifing.

We modeled both hybridization and incomplete coalescence under the coalescent network framework implemented in Phylonet v3.5 (*-InferNetwork_MPL*). Networks were computed under the Maximum Pseudo-Likelihood (MPL) criterion from the 6724 *Agraulis*-rooted autosomal phylogenies,

considering only nodes with support >0.8 (*-b 0.8*) and starting with the MP-EST species. The analysis was conducted separately for each clade among *Eueides, H. melpomene, H. erato, H. sara, H. hecuba* and *H. egeria*. To detect ancient admixtures between basal branches of the tree, we conducted an analysis where all individuals of each clade were treated as a single "species". The optimal network was determined by calculating the BIC from the Maximum Likelihood and the number of lineages, admixture edges and gene trees in each model.

Heliconius hecalesia and *H. clysonymus* were difficult to place with confidence in the phylogenies, and clustered with *H. hortense* at the *B/D* locus. To test for gene flow between these species and relatives we calculated the *D* statistic (Durand et al. 2011), determining significance by block jacknifing (Martin et al. 2015). Specific tests were conducted for gene flow between *H. hecalesia* (recipient, P2) and *H. clysonymus, H. hortense, H. telesiphe* or species from the *H. sara* clade (donor, P3). The sister species P1 was either the allopatric *H. erato* from French Guiana, or the parapatric *H. erato* from Amazonia, and the outgroup was always *H. melpomene*. To test the possibility that the lack of resolution at one node in the larger phylogeny (Fig. 1; S1 Figure) is caused by ancient admixture prior to the formation of current species, we conducted tests treating entire clades as taxa (Table 3). The extent of similarity between species clusters was illustrated with PCAs of genome-wide variation, calculated for the *H. erato/sara* clade in the R package *adegenet*(Jombart & Ahmed 2011). A chronogram for the putatively introgressing *B/D* red pattern intervals (HE670865: 320-380 and 410-440 kbp) was estimated with RelTime.

Introgression at the colour pattern loci

Out of 4309 scaffolds, the five fragments containing introgressing loci associated with aposematic wing phenotypes (Table 2) were treated separately. Each scaffold alignment was partitioned into windows of 20 kbp, sliding by 10 kbp, discarding windows with less than 1000 bp of data [script sliPhy3.py(Martin 2017)]. The topology for every window was reconstructed with FastTree, tested for significant differences from the MP-EST species tree using the SH test (Shimodaira & Hasegawa

1989) in RAxML and inspected visually. In order to understand precisely how the *Hmel1* reference corresponds to the specific red control loci of *H. erato* (Supple et al. 2013), the *B/D* scaffold HE670865 was aligned against the *H. erato B/D* BACs (Papa et al. 2008) with mLAGAN (Brudno et al. 2003).

ABBREVIATIONS

ARG, Ancestral Recombination Graph; BIC, Bayesian Information Criterion; IC, Internode Certainty; ILS, Incomplete Lineage Sorting; LD, Linkage Disequilibrium; ML, maximum likelihood; MSC, Multispecies Coalescent; SNP, single nucleotide polymorphism

ACKNOWLEDGMENTS

We thank the governments of Peru, Ecuador and Suriname for permits to collect butterflies. Kanchon Dasmahapatra, James Mallet and Camilo Salazar provided advance access to genomic data and helpful comments. Analyses were conducted on a machine made available by Aylwyn Scally, clusters at the School of Life Sciences operated with help from Jenny Barna, and the STRI *Plato* server run by Eugenio Valdes. Richard Nichols and John Welch examined an early draft of the manuscript.

FUNDING

KMK was funded by Herchel Smith, Balfour-Browne and Cambridge Philosophical Society studentships. Funding for a collecting trip to Suriname was provided by the Panton Trust of Emmanuel College. WOM was funded by the Smithsonian Institution. MJ was funded by French ANR Grant HYBEVOL (ANR-12-JSV7-0005) and research permits 288-2009-AGDGFFS-DGEFFS and 0148-2011-AG-DGFFS-DGEFFS from the Peruvian Ministry of Agriculture. CJ was funded by a European Research Council Grant 339873.

MAIN TABLES AND FIGURES

Recipient	*B/D* Donor	*Yb/Cr*	*Ac/Sd*	*Ro*	*K*	frequency	Autosomes-wide
H. hecalesia	*H. clysonymus/ H. hortense*	*H. clysonymus/ H. hortense*	*H. clysonymus/ H. hortense*	*H. clysonymus/ H. hortense*		0.140	Variation shared with *H. clysonymus, H. sapho* and *H. erato* clades.
H. numata						0.003	*H. melpomene* E
H. hecale clearei	*H. pardalinus; H. numata; H. ethilla*					0.046; 0.005; 0.019	
H. elevatus	*H. melpomene* E		*H. melpomene* E, *H. cydno*			0.003; 0.001	
H. melpomene E	*H. elevatus*					0.003	
H. pardalinus/elevatus		*H. melpomene/ H. cydno*				0.000; 0.004	*H. melpomene* E
H. timareta/heurippa	*H. melpomene* E	*H. melpomene* E		*H. melpomene* E	*H. melpomene* E	0.043	*H. melpomene* E
H. timareta			*H. melpomene* E			0.090	
H. heurippa	*H. melpomene* W		*H. cydno/H. pachinus*			0.008; 0.043	
H. cydno/timareta		*H. melpomene* W/FG				0.000	
H. cydno/pachinus		*H. melpomene* W				0.043	*H. melpomene* W
H. m. malleti Colombia			*H. cydno/H, timareta*			0.000	
H. pardalinus/elevatus/ hecale/atthis/ethilla						0.008	*H. melpomene* E

Table 1. Loci and direction of introgression varies between species. An overview of incongruences at the colour pattern loci detected by inspecting ML gene trees, and genome-wide admixture found with TreeMix and MPL networks. Frequency of the clusters counted among autosomal gene trees. Novel findings highlighted in yellow.

H. melpomene	***H. erato***	***Other***	**Genes**	**Scaffold**	**Phenotype**	**Key references**
B	*D*	*Br/G*[1]	*optix, kinesin,* putative enhancers	HE670865	Red on HW and FW, ventral brown patterns	Reed et al. 2011; Supple *et al*., 2013; Wallbank et al. 2016; Van Belleghem et al., 2017; Zhang et al. 2017
Yb/Sb/N	*Cr*	*P*[2]	*Cortex,* putative enhancers	HE667780	Yellow/white on HW and FW	Nadeau et al. 2016; van Belleghem et al., 2017; Enciso et al., 2017
Ac	*Sd*		*wntA,* putative enhancers	HE668478 HE669520	Pattern shape	Martin et al. 2012; Mazo-Vargas et al. 2017
Ro	*Ro*		*vvl, rsp3*	HE671554	FW band shape	Nadeau et al. 2014; Van Belleghem et al., 2017
K	*K*		*wingless*	HE671246 HE670889	white/yellow switch	Kronforst et al., 2016

Table 2. Major wing pattern and colour loci in *Heliconius*. The key loci are named differently in divergent species (Sheppard et al. 1985), but in all cases the approximate genomic location has been identified, good candidate genes have been identified and intervals containing functional variation have been localized (Reed et al., 2011; Martin et al., 2012; Nadeau et al. 2016; Wallbank et al., 2016; Van Belleghem et al. 2017). Scaffold numbers refer to the *Hmel v1* assembly. [1]Brown patterns in *H. cydno* and *H. pachinus* (Chamberlain et al. 2011); [2]The *Pushmipullyu* supergene controlling most of the wing patterning in *H. numata* (Joron et al. 2011). HW: hindwing; FW: forewing.

P1	*P2*	*P3*	*D*	error(*D*)	*p*-value
erato FG	*hecalesia*	*clysonymus*	0.349	0.009	<0.0001
erato FG	*hecalesia*	*hortense*	0.383	0.009	<0.0001
erato FG	*hecalesia*	*telesiphe*	0.269	0.008	<0.0001
erato FG	*hecalesia*	*charithonia+peruvianus*	0.208	0.005	<0.0001
erato FG	*hecalesia*	*sara+leucadia*	0.170	0.005	<0.0001
erato East	*hecalesia*	*clysonymus*	0.354	0.009	<0.0001
erato East	*hecalesia*	*sara+leucadia*	0.178	0.004	<0.0001
erato clade	*hecalesia*	*clysonymus* clade	0.305	0.008	<0.0001
erato clade	*hecalesia*	*sara* clade	0.171	0.004	<0.0001
hecalesia	*erato* clade	*clysonymus* clade	-0.305	0.008	*NA*
hecalesia	*erato* clade	*sara* clade	-0.171	0.004	*NA*

Table 3. D-statistic values (ABBA/BABA tests) of admixture between *H. hecalesia* and relatives. Lower section of the table describes tests treating entire clades as taxa, to detect ancient admixture. P2 is the hypothetical recipient and P3 the donor of variants. Positive *D* values are evidence for admixture after accounting for ILS. *p*-values calculated by block jacknifing.

FIGURES

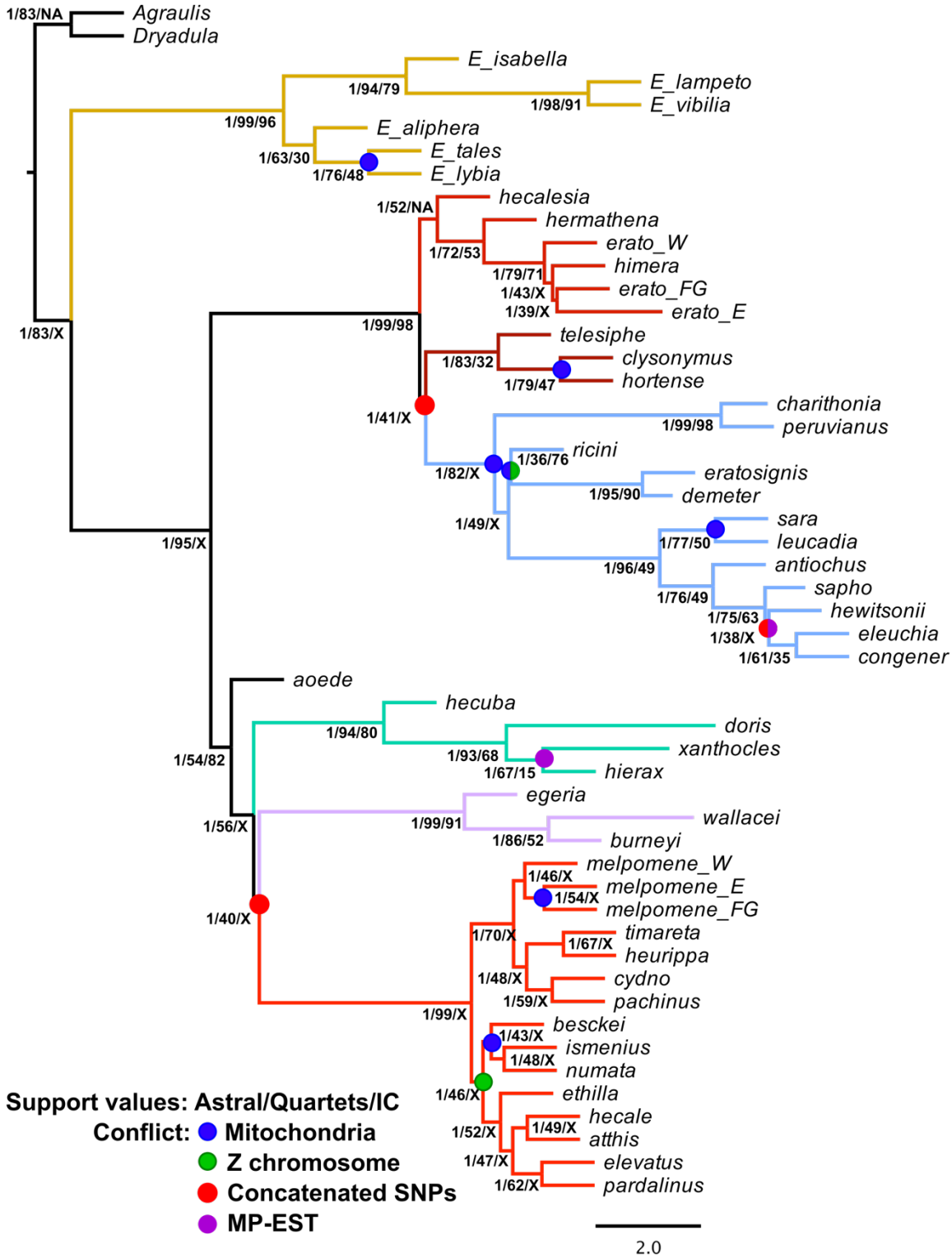
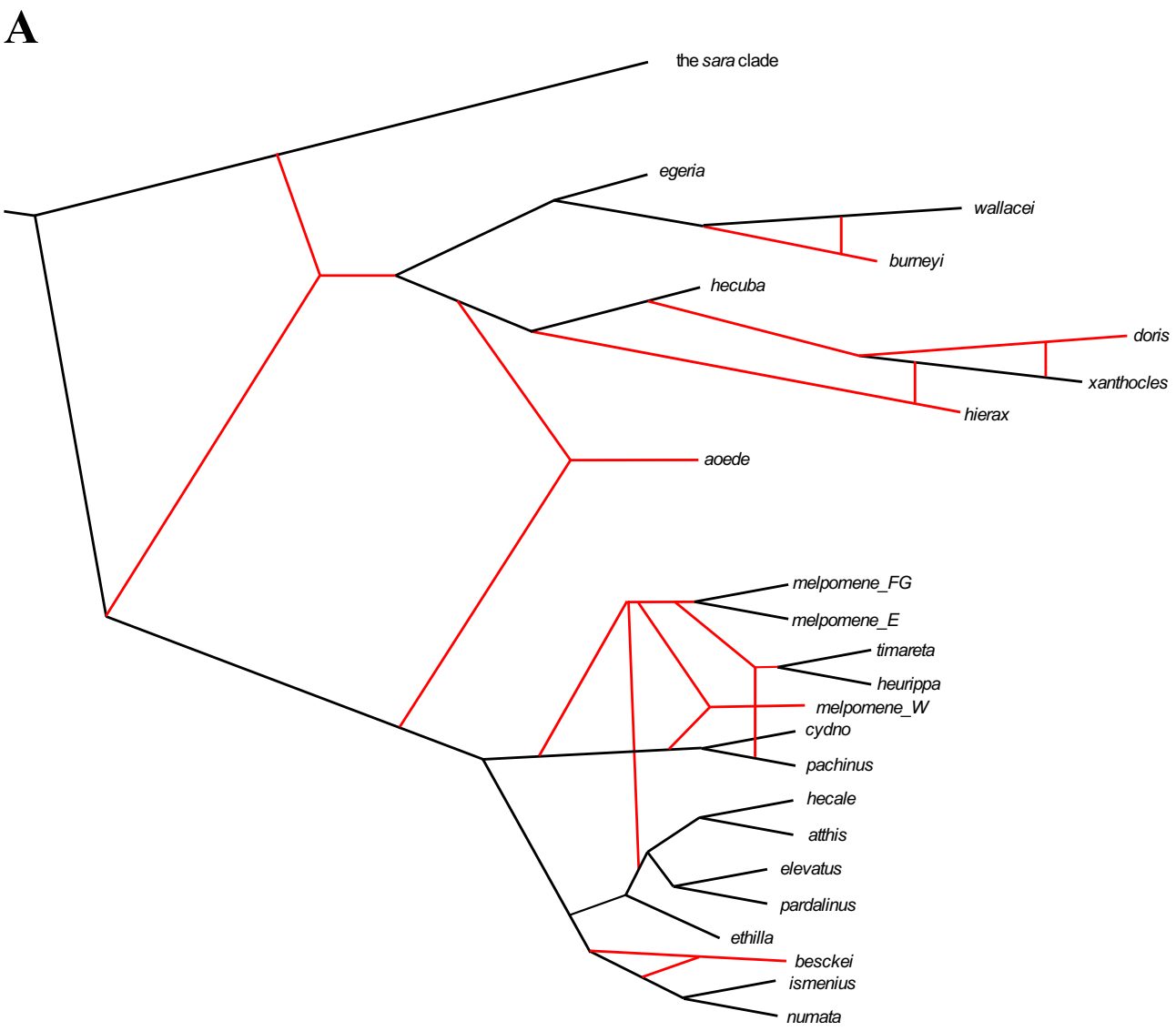


Figure 1. Well-resolved phylogeny of Heliconiini obscures the underlying incongruence. The topology identified by the multispecies coalescent method ASTRAL is well supported, although many of the nodes are not represented in several of the gene trees (numerical values). Topologies are consistent between coalescent and concatenation methods based on whole genome data, but conflicted mainly with the mitochondrial phylogenies (blue dots). Branch colours correspond to previously defined clades (Brown 1981; Kozak et al. 2015): red – *melpomene*/Silvaniforms; violet – *egeria*; green – *hecuba*; blue – *sara*; crimson – *clysonymus*; scarlet – *erato*; brown – *Eueides*.



B

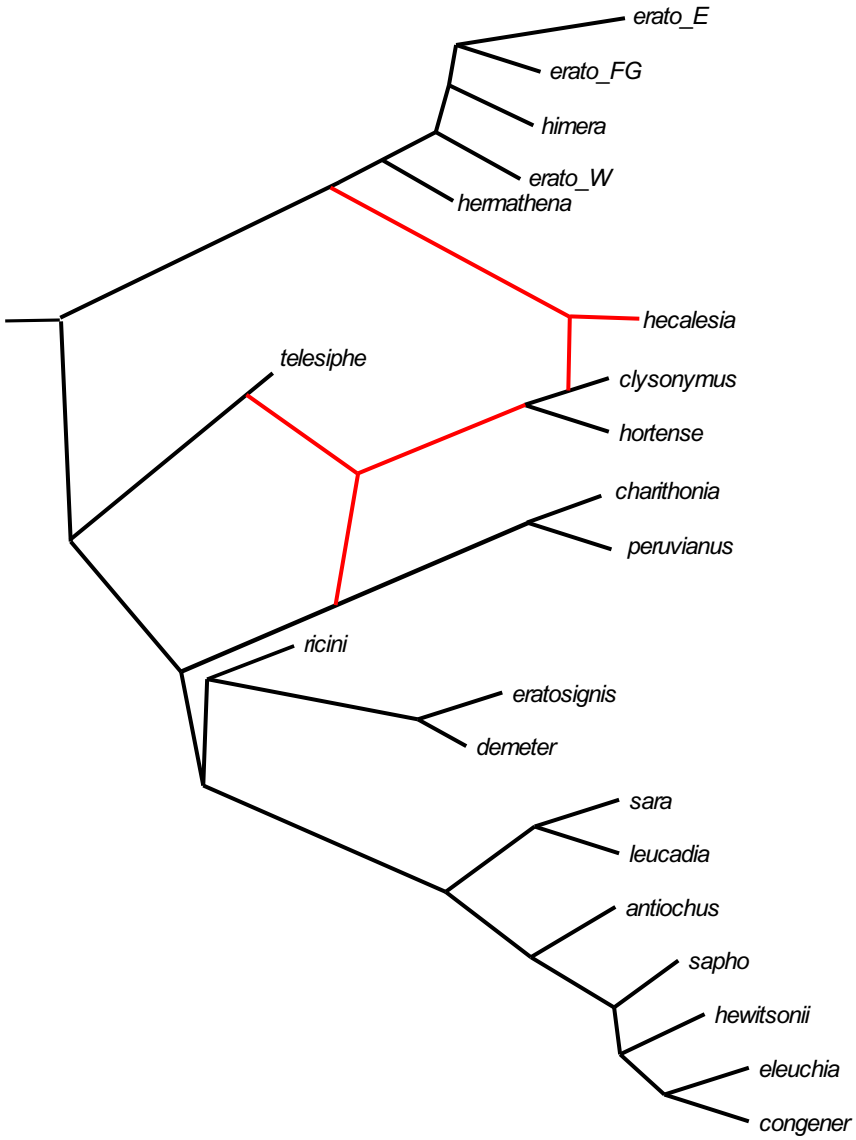


Figure 2. Evidence of introgression is found across the entire *Heliconius* radiation. Networks inferred under Maximum-Pseudolikelihood (MPL) based on 6724 autosomal ML gene trees distinguish between introgression and incomplete lineage sorting, revealing previously unknown admixture events (red edges). (A) The analysis of *H. melpomene* and cognates reveals several unknown introgressions closer to the root of the tree. Known events are recapitulated, demonstrating robustness of the approach. (B) Relatively few admixtures occurred in the evolution of the *Heliconius erato/sara* clade, but *H. hecalesia* may be a previously unknown recent hybrid.

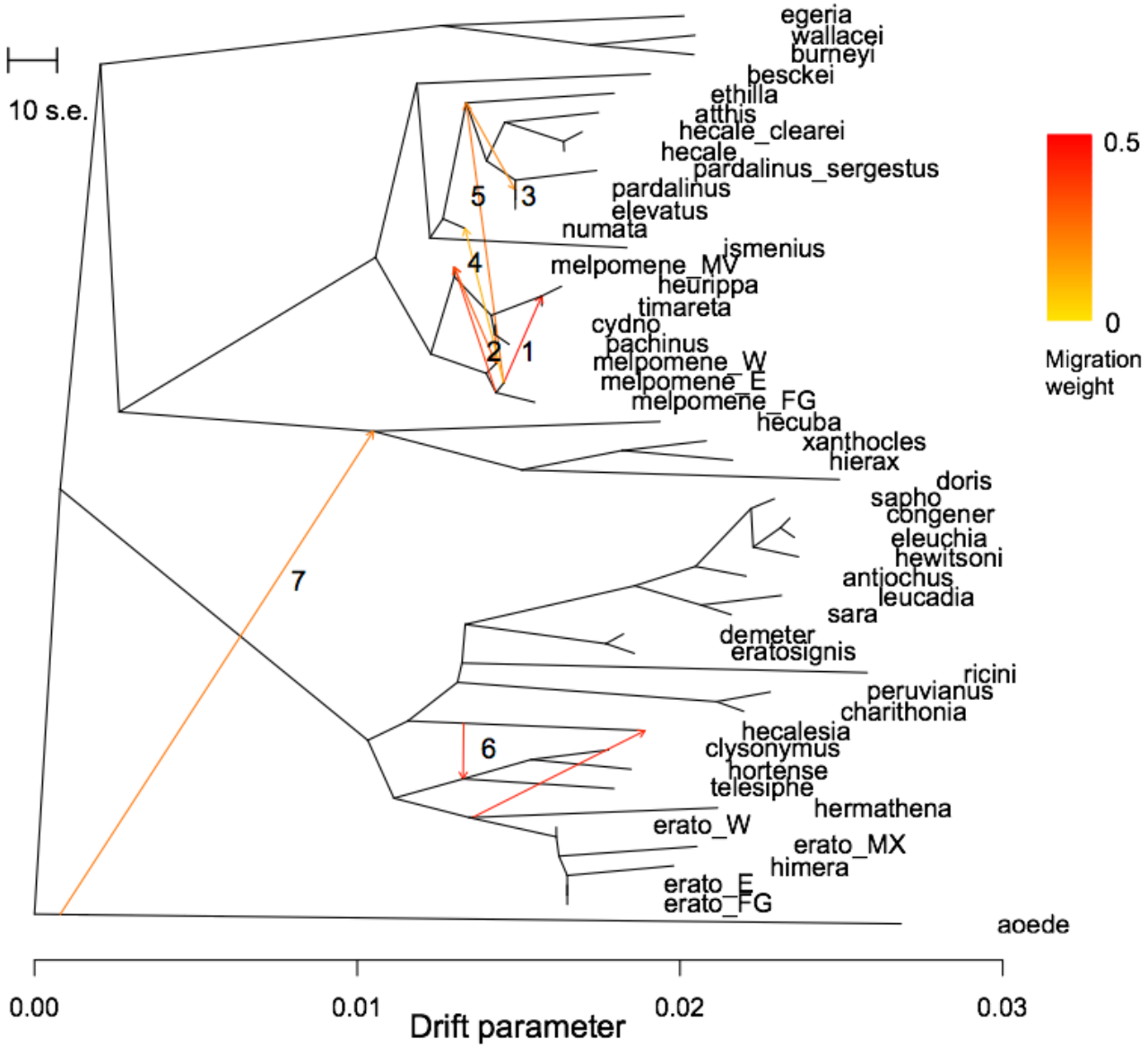


Figure 3. The extent of interspecific gene flow varies across the tree. TreeMix inference of splits and mixture from autosomal SNPs. Migration edges (1-7) are inferred on a phylogenetic tree built from allele frequencies under a Gaussian genetic drift approximation. Colours of the edges correspond to the proportion of the genome exchanged.

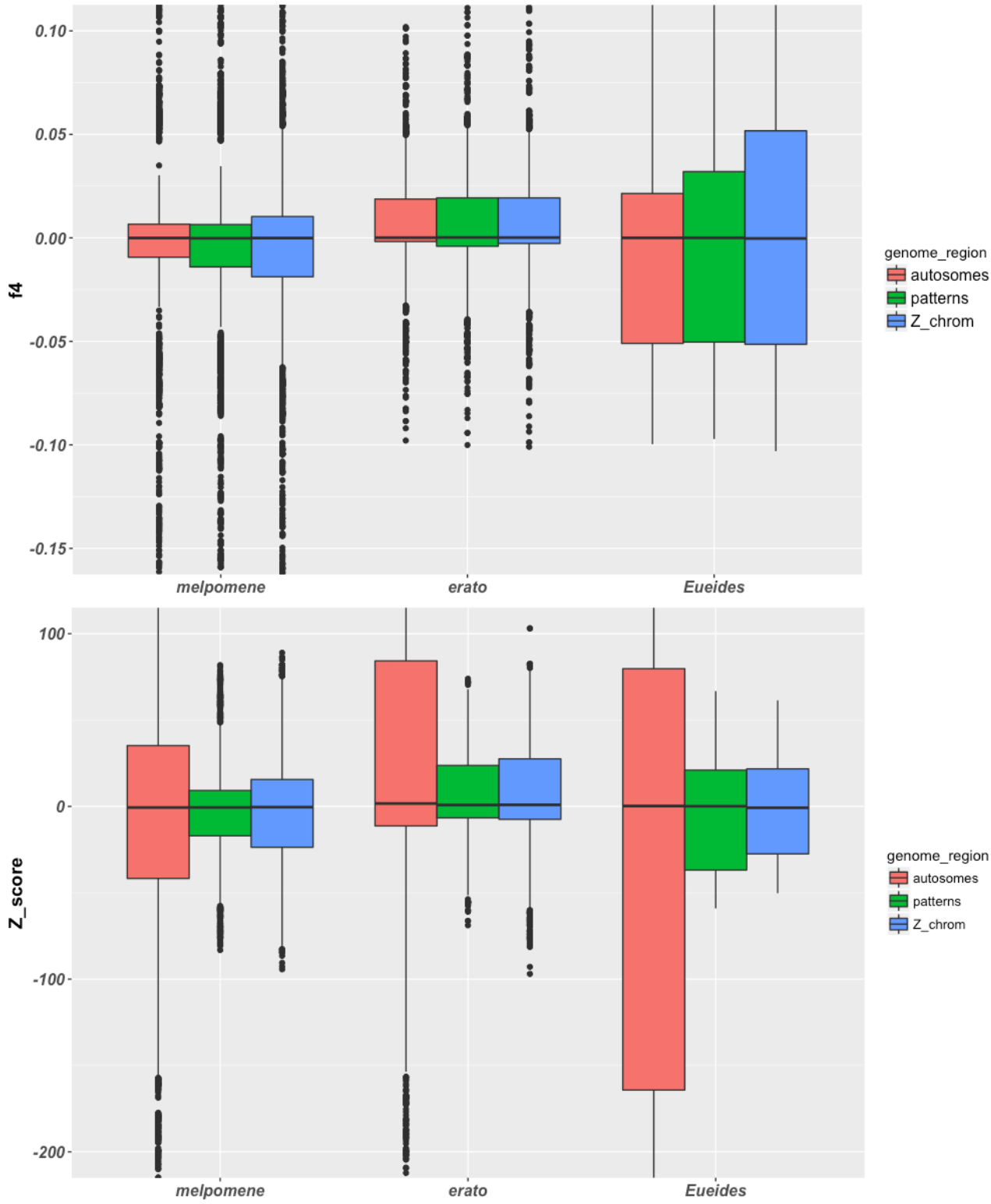


Figure 4. Structure of the *Heliconius erato* clade is the closest to the bifurcating tree model.

(A) When introgression among all possible quartets is tested in the major clades, more admixture events (negative *f4* statistics) are found in the *H. melpomene* than *H. erato* group. *Eueides* shows

high variation possibly due to incomplete sampling. Gene flow is more common at putatively neutral autosomal sites than adaptive patterning or sex-linked loci. (B) *Z*-scores for all possible *f4* estimates in each clade calculated by jackknifing . Values below -4.47 indicate a significant test at the 0.01 level after correction for multiple testing.

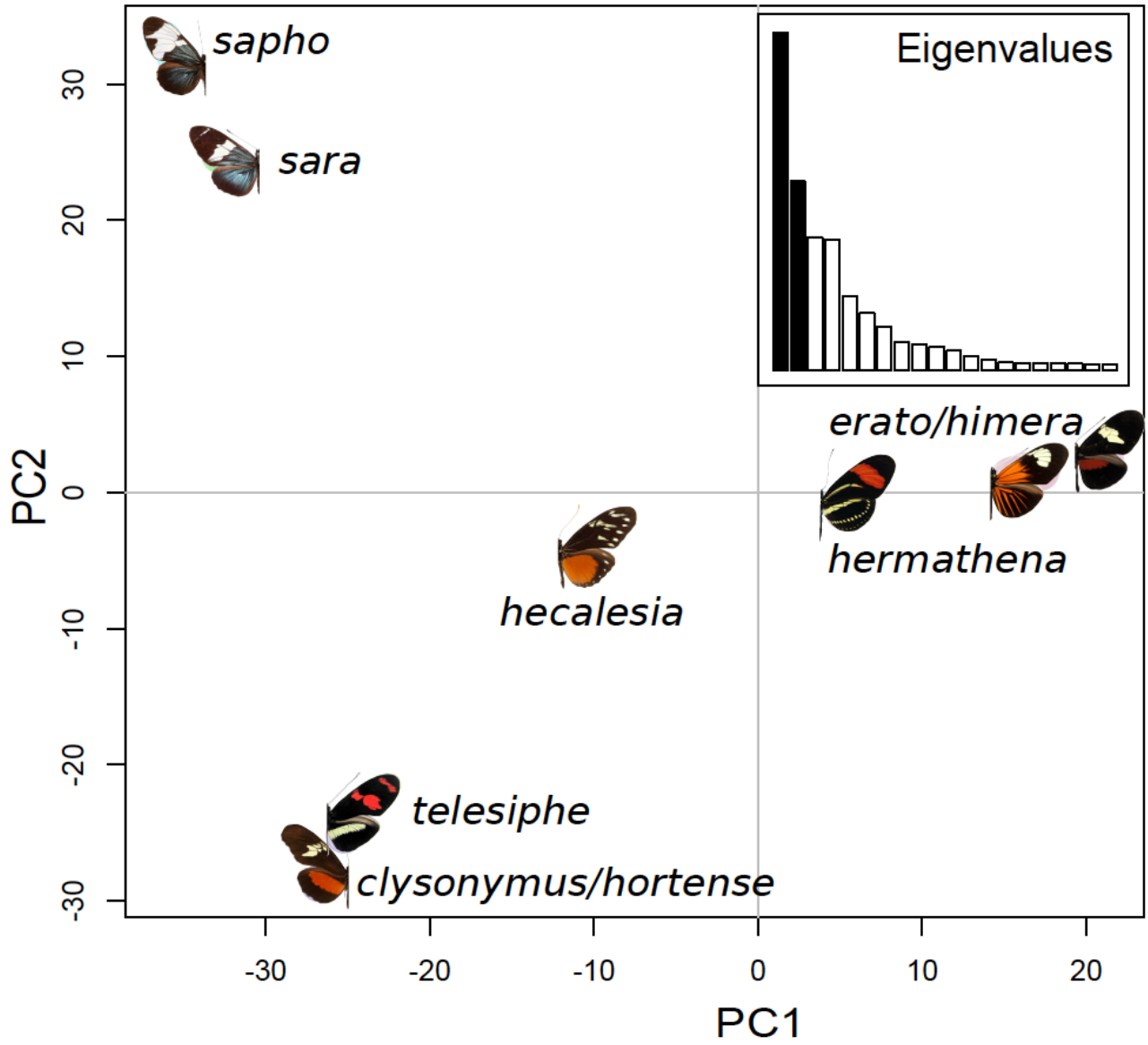


Figure 5. Ambiguous genomic composition of *H. hecalesia*. Although usually recovered as the sister species of *H. hermathena* and *H. erato* in bifurcating phylogenies, H. hecalesia shares substantial amount of variation with two other clades. Principal Component Analysis of variation in the autosomal exonic SNPs within the *H. erato/sapho* clade. First two PCs plotted, accounting for over half of the variation.

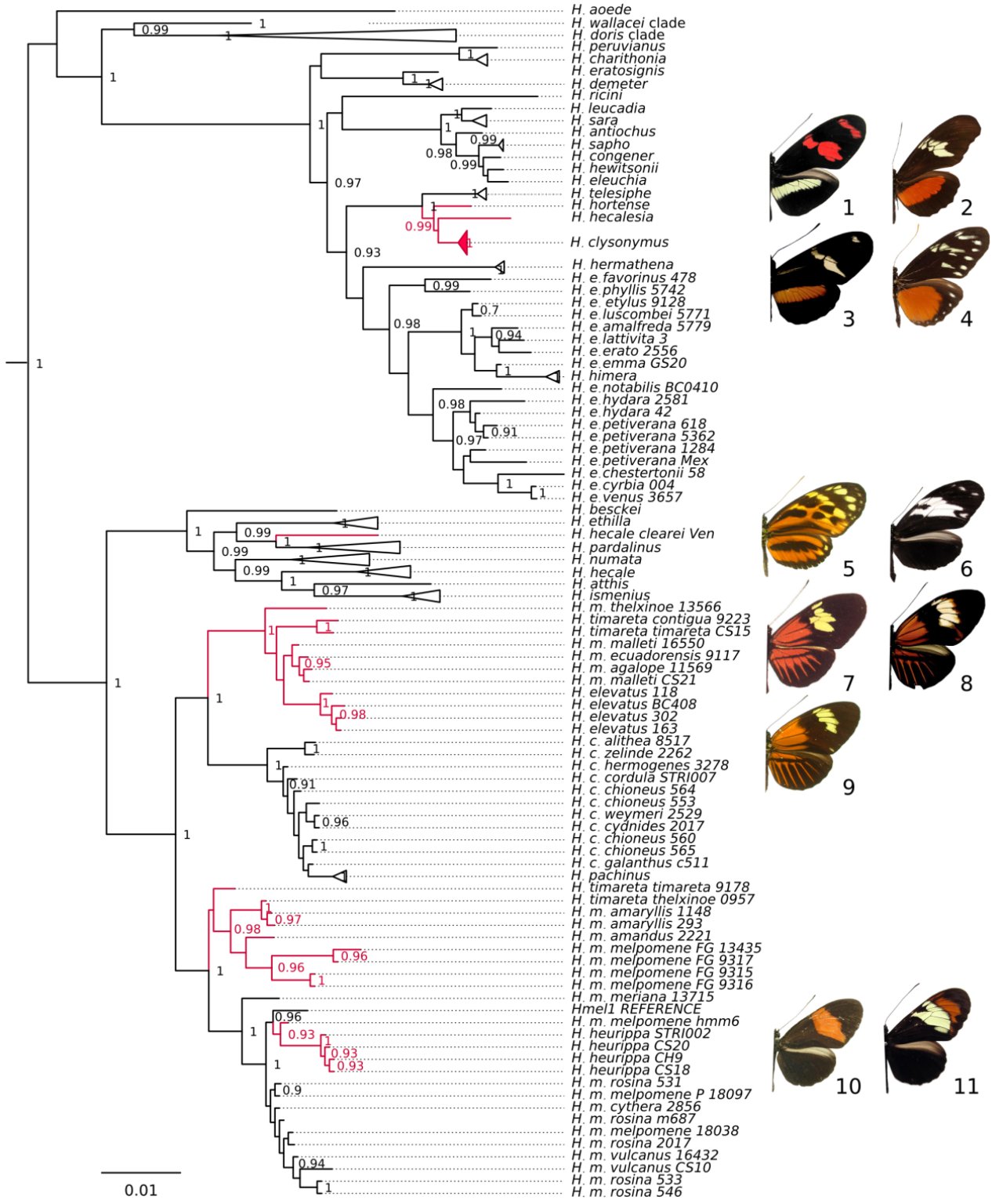


Figure 6. Pervasive introgression across the species boundary at the red patterning locus.

Branches in unexpected positions labeled red. The ML tree was reconstructed for the 360,000-380,000 bp region on the B/D scaffold, including the main peaks of association with patterns in *H. erato* and *H. melpomene*. Intraspecific relations for species not discussed in text are collapsed.

Outgroups and parametric support values <0.9 not shown. 1. *H. telesiphe sotericus,* 2. *H. hortense*, 3. *H. clysonymus hygiana*, 4. *H. hecalesia formosus*; 5. *H. hecale felix*, 6. *H. hecale clearei*; 7. *H. timareta timareta*, 8. *H. melpomene malleti*, 9. *H. elevatus*; 10. *H. melpomene melpomene* (Magdalena Valley), 11. *H. heurippa*.

SUPPLEMENTARY METHODS

Software command lines and parameters

Mapping:

Settings for the most divergent genomes were based on experiments with *Eueides lybia* and *Acraea encedon* libraries (BWA mismatch numbers k={2, 3}; l ={15-35}; Stampy substitution rate ={0.1-0.2}). The proportion of properly paired reads did not change by more than 4% among the combinations, and most conservative values were used as listed in S2 Table.

GATK UnifiedGenotyper calling:

GenomeAnalysisTK.jar -T UnifiedGenotyper -R Hmel1.1.fasta -I individual.bam -o individual_variants.vcf --output_mode EMIT_ALL_SITES --downsample_to_coverage=250 --baq=CALCULATE_AS_NECESSARY

TrimAl v1.2 alignment trimming for of sites with most data missing:

trimal -in alignment.fasta -out alignment.trimal.fasta -fasta -resoverlap 0.5 -seqoverlap 50

Removal of uninformative sites by Block Mapping and Gathering with Enthropy (BMGE):

BMGE.jar -i alignment.trimal.fasta -t DNA -m DNAPAM100:1 -ory alignment.RY.phy -s NO -g 0.5

FastTree v2 ML gene tree estimation:

FastTree -nt -gtr -pseudo -spr 4 -gamma -log alignment.ft.log $f > alignment.ft.tre

RAxML v8 ML supermatrix tree estimation:

raxmlHPC-PTHREADS-SSE3 -T 10 -f d -p 123 -m GTRGAMMA -s supermatrix.fasta -n supermatrix.ML.tre

RAXML v8 Internode Certainty calculation:

raxmlHPC-PTHREADS-AVX –T 10 -f i -m GTRCAT –t supermatrix.tre -z gene.trees -n ica.analysis

MP-EST v1.4 control file lines:

gene.trees

1

-1

6848 52

{mapping of individuals to species}

0

Astral-III command:

astral.5.5.9.jar -i in.tre

astral.5.5.9.jar –i gene.trees –t 1 –a mapping.to.species.clust –o astral.tre

Plink v2 SNP filtering:

plink --file treemix.plink –ld –freq –noweb –missing 0.05 --within mapping.to.species.clust

TreeMix v1.13 ARG with one admixture:

treemix -i treemix.calls.gz –m 1 –k 100 -root H_erato -o H_melpomene_clade

TreeMix v1.13 Reich's f_4 statistics:

fourpop -i treemix.calls.gz –m 1 –k 100 > f4.out

Phylonet Maximum Pseudo-Likelihood network with one hybridisation event:

InferNetwork_MPL (Autosomal_tree_1-Autosomal_tree_6725) 1 -n 3 -pl 12 -x 10 -s

mpest_start_tree -di -b 0.8

Results of read mapping across the genus to the *Hmel v1* reference

New Whole Genome Illumina resequencing data were produced successfully for 11 species, extending the sampling of the *H. sara* clade and *Eueides*. For all Illumina samples, the mean per-base quality was high along the entire reads (Q>28) and all the quality filters were passed. The number of reads ranged from 94,857,214 to 194,528,700 per individual, corresponding to an expected coverage of 34x to 71x (S3 Table).

Most clades within *Heliconius* (as defined in Brown 1981 and Kozak et al. 2015) were sampled thoroughly, except the inclusion of only one out of four species in the subgenus *Neruda* (see Chapter 2 for taxonomy). Among the 23 *Heliconius* species with multiple samples, the number of individuals ranged from two to 25, the number of geographically distinct populations was between one and 12, and the number of colour pattern races ranged from one to 14 (S1 Table).

The depth and number of mapped reads decreased predictably with increasing divergence from the reference (Table 5.3). In case of the MCS clade, more than 90% of reads mapped to the *Hmel1*, and most were properly paired with their mate. The number of mapped reads fell steeply for more distantly related taxa, averaging 58.6% for *H. erato,* 43.2% for *Eueides* and around 30% for the outgroups. As the protein coding sequence comprises 5% of the *H. melpomene* genome (Heliconius Genome Consortium 2012), the data obtained for these more distantly related species is mostly non-coding sequence. The non-heliconian outgroup *Acraea encedon* (Acraeini) and the *H. doris* sample 8684 were excluded due to low effective coverage (respectively 1.5x and 3.4x) manifested in poor quality of inspected alignments. Regardless of divergence from the reference the mapping was the highest for samples with expected coverage >40x and usually low at <15x.

The number of SNPs follows similar trends, The highest number of credible variants was called in the MCS clade (S3 Table). More distant taxa showed naturally higher variation from the reference, which simultaneously reduced the overall number of mapped sites. In total, 126,865,683 SNPs were identified, a number driven primarily by taxon-specific differences from the reference,

but also by high variability in the densely sampled MCS group. For instance, private SNPs constituted 30% of the total and $11.5*10^6$ private variants were discovered just in the *H. melpomene* samples. 5,483,419 SNPs were found in the exome. Overall, these findings are consistent with previous reports for the MCS group (Martin et al. 2013) and the monarch butterflies *Danaus* (Zhan et al. 2014).

The transition/transversion ratio for the MCS clade is a low 1.25, but similarly small values were previously found in the noncoding sequences of the cricket *Podisma pedestris* (Keller et al. 2007). The bias increases with divergence, reaching 1.43 for *Eueides* and 1.50 for *Dryadula*, most likely due to a higher proportion of CDS in the recovered total (Table 5.3). Such variation in the Ts/Tv bias is observed in the human data, where the Ts/Tv equals 2.1, but increases to 3.0 in the exome (1000 Genomes Initiative 2015).

SUPPLEMENTARY TABLES

S1 Table. Specimens – XLS file.

Clade	Divergence (Myr)	BWA *k*	BWA *l*	Stampy *subRate*
H. melpomene	<1.5	2	32	0.03
H. cydno	2.0	2	32	0.04
Silvaniforms	4.0	2	32	0.05
H. erato	12.0	2	25	0.1
Eueides, Agraulis, *Acraea*	>18.5	2	25	0.1

S2 Table . Empirically adjusted parameters for the short read alignment to the *H. melpomene* reference. *k*=maximum number of mismatches per 100 bp; *l*=minimum stretch of identical sequence necessary to map; *subRate*=expected nucleotide divergence.

Clade	Species	Samples	Coverage x	Reads mapped*	Reads properly paired*	# SNPs	Biallelic SNPs	Singletons	Ts/Tv
H. melpomene	1	25	27.33 (5.19-85.58)	316,265,364 (94.55%)	262,954,142 (78.62%)	32,788,205	30,615,977	11,497,421	1.27
H. melpomene/ H. cydno	5	48	26.01 (5.19-100.01)	388,402,453 (93.80%)	309,082,266 (74.64%)	48,793,385	44,279,711	16,910,938	1.26
Silvaniform (*H. numata*+relatives)	8	28	21.82 (8.05-33.69)	142,606,506 (90.21%)	97201540 (61.49%)	57,561,620	50,888,801	21,259,708	1.24
H. wallacei	3	4	10.16 (5.32-16.80)	81,888,086 (67.03%)	31084824 (25.45%)	17,425,503	16,832,134	3,341,464	1.28
H. doris	4	6	8.46 (3.43-22.85)	172,414,557 (72.19%)	70,011,956 (29.31%)	35,334,399	32,579,459	5,613,625	1.25
H. aoede	1	1	14.36	79,473,860 (61.97%)	31,150,826 (24.29%)	8,581,604	7,257,197	8,581,604	1.27
H. erato	7	33	10.57 (5.11-16.55)	138,223,134 (64.01%)	49,770,828 (23.05%)	33,988,575	30,643,397	7,578,237	1.32
H. sara	12	17	10.02 (5.55-19.07)	138064019 (58.63%)	48,057,198 (20.41%)	26,791,561	24,831,585	4,599,643	1.30
Eueides	6	6	6.07 (4.90-7.54)	72,571,204 (43.22%)	15,904,356 (9.47%)	12,905,223	12,005,724	1,966,268	1.43
Agraulis	1	1	7.26	58,461,652 (30.57%)	14,577,290 (7.62%)	4,949,437	4,459,801	4,949,437	1.47
Dryadula	1	1	3.61	27,336,635 (30.85%)	6,130,764 (6.92%)	4,141,846	3,975,796	4,141,846	1.50
TOTAL	**48**	**145**	**17.4 (3.43-100.01)**	**n/a**	**n/a**	**126,865,683**	**90,646,525**	**38,070,723**	**1.29**

S3 Table. Mapping quality and number of SNPs decrease with divergence from the reference. Statistics for the BWA/Stampy read mapping of Illumina 100 bp paired-end reads to the *H. melpomene* reference, averaged by clade *sensu* Brown 1981. Values were calculated for sites with quality score 20 or higher. Ranges reported in parentheses. *Percentages of reads mapped reported for the best sample.

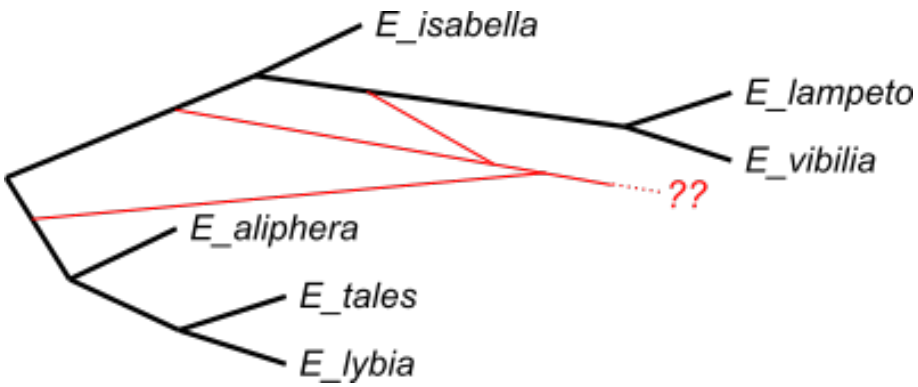
S4 Table. TreeMix and Phylonet Model selection –XLS file.

Parameter	Autosomal	Z-linked
# alignments	6848	416
# *Agraulis*-rooted alignments	6724	406
Taxa after TrimAl	144.53	144.75
Length before BMGE (bp)	1399	1633
Length after BMGE (bp)	1387 (60-15,921)	1627 (210-11,979)
Missing data	4.0 %	3.6 %
Ambiguous sites	1.3 %	0.3 %
GC content	42.9 %	44.6 %
Interspecific pairwise identity	90.8 %	89.7 %
Gene tree length	0.7128	0.7416

S5 Table. Basic statistics for the autosomal and Z-linked protein-coding gene alignments.

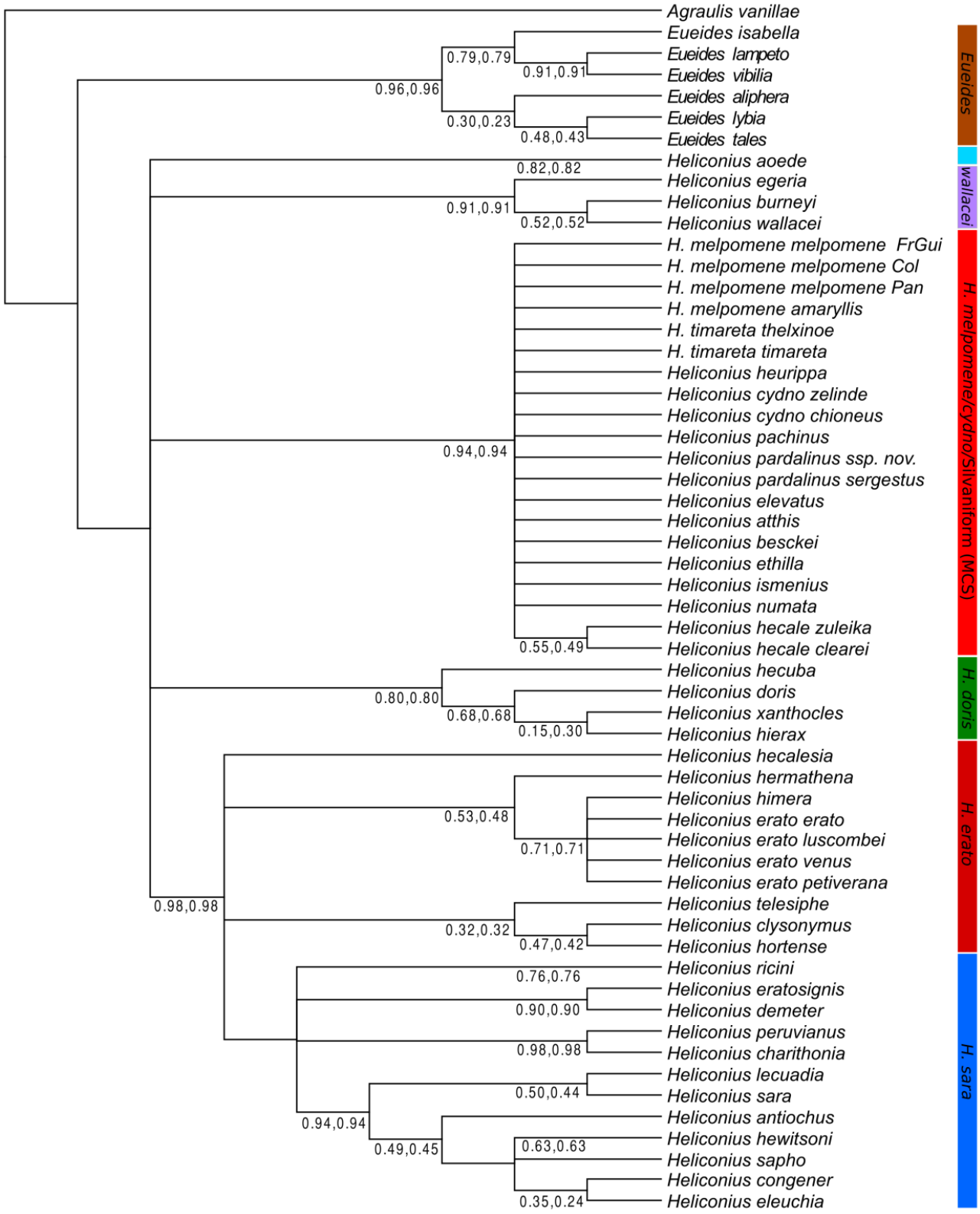
Relatively short sequences were removed with TrimAl and uninformative sites were deleted by Block Mapping and Gathering with Entropy (BMGE). Range of lengths after trimming in parentheses.

SUPPLEMENTARY FIGURES



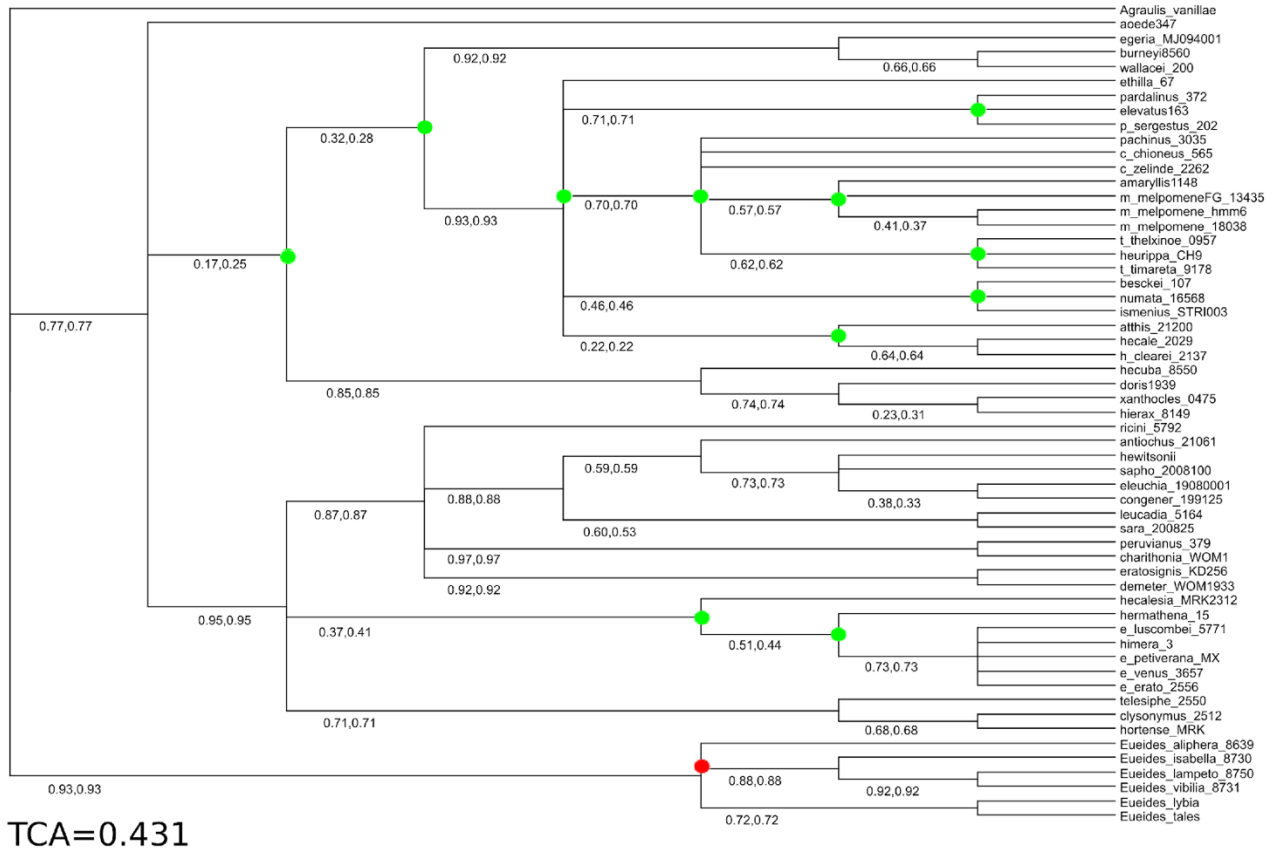
S1 Figure. Inconsistent estimates of gene flow in a representation of *Eueides* (6/12 species), the sister group of *Heliconius*. MPL network from the autosomal gene trees. The vertices of multiple branches may represent an artefactual "ghost lineage" corresponding to missing species.

S2 Figure. **High support for a concatenation tree based on autosomal SNPs.** All nodes in the Maximum Likelihood (RAxML) phylogeny have a bootstrap support of 100, except for the split labeled with a red dot (62/100). Most intraspecific samples collapsed. Branches coloured by clade as defined in Chapter 2: brown – *Eueides*; red – *H. sapho* clade; navy – *H. erato* clade; blue – *H. aoede* clade (formerly *Neruda*); green *H. doris* clade; violet – *H. wallacei* clade; orange – Silvaniforms; red – *H. cydno* and cognates; pink – *H. melpomene*.

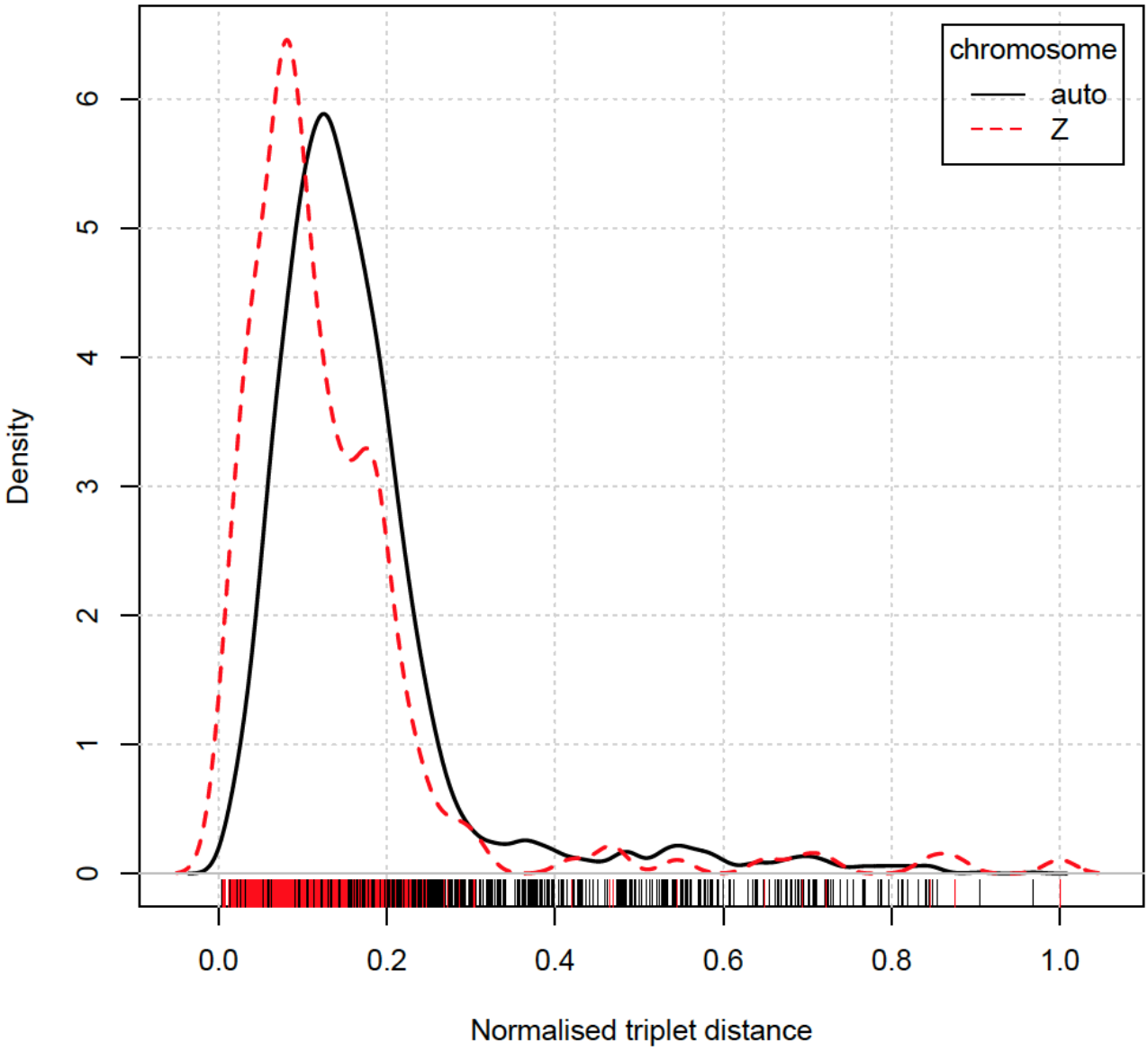


S3 Figure. Autosomal gene trees disagree at most nodes. 50% Majority Rule Consensus tree based on 6724 Agraulis-rooted gene trees. Branch labels indicate the IC and ICA support (Salichos

et al. 2014). An IC=0.0 means that a given node has an equally frequent alternative in the distribution of gene trees, whereas IC=1.0 means that all trees contains this node. FrGui: French Guiana; Col: Colombia; Pan: Panama.

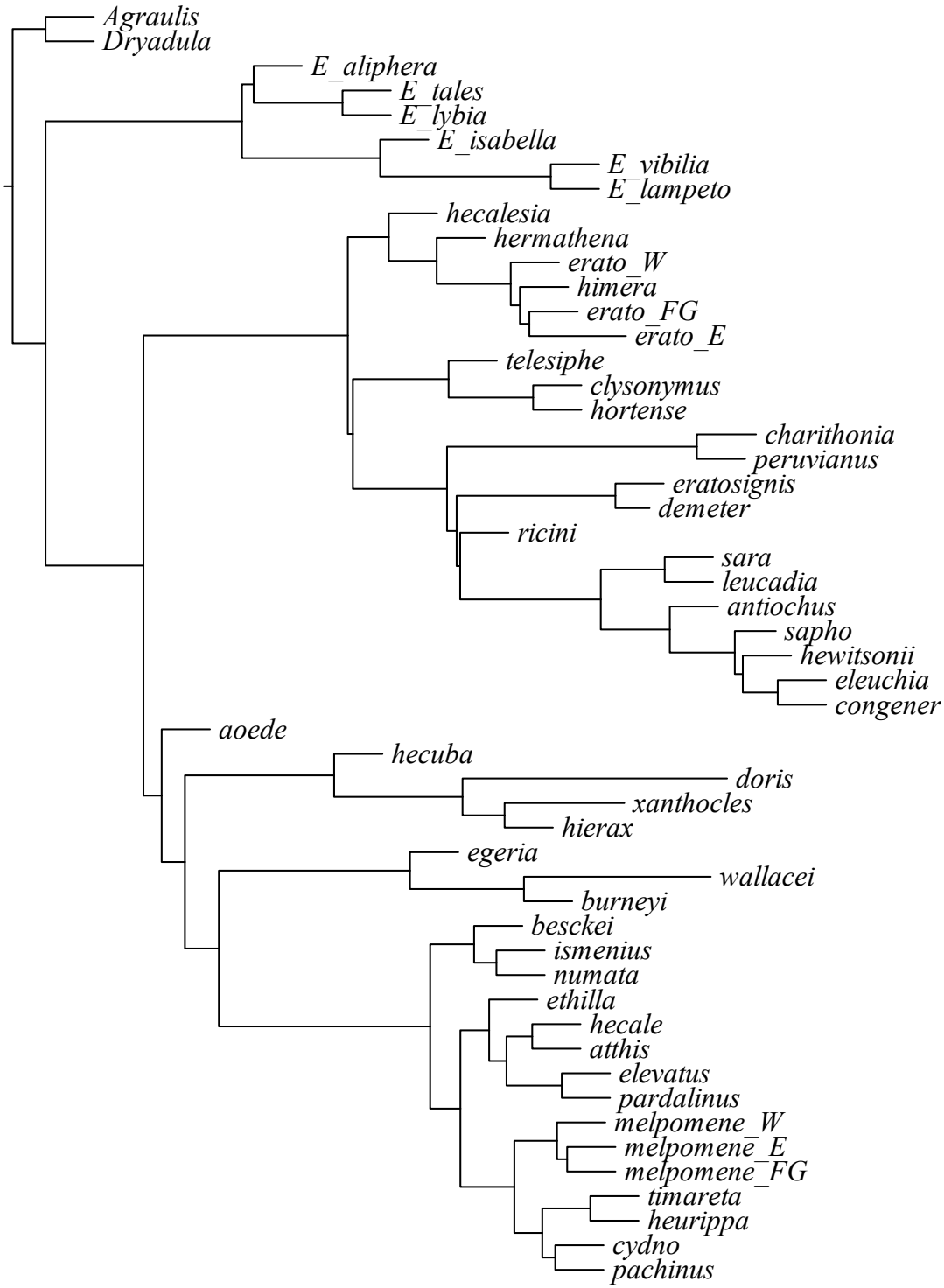


S4 Figure. More interspecific genetic isolation at the Z chromosome than at the autosomes. 50% Majority Rule Consensus the Z-linked gene trees with IC/ICA support values indicated. Green dots indicate nodes unresolved in the autosomal 50% MRC, red dots nodes conflicting with the autosomal tree.

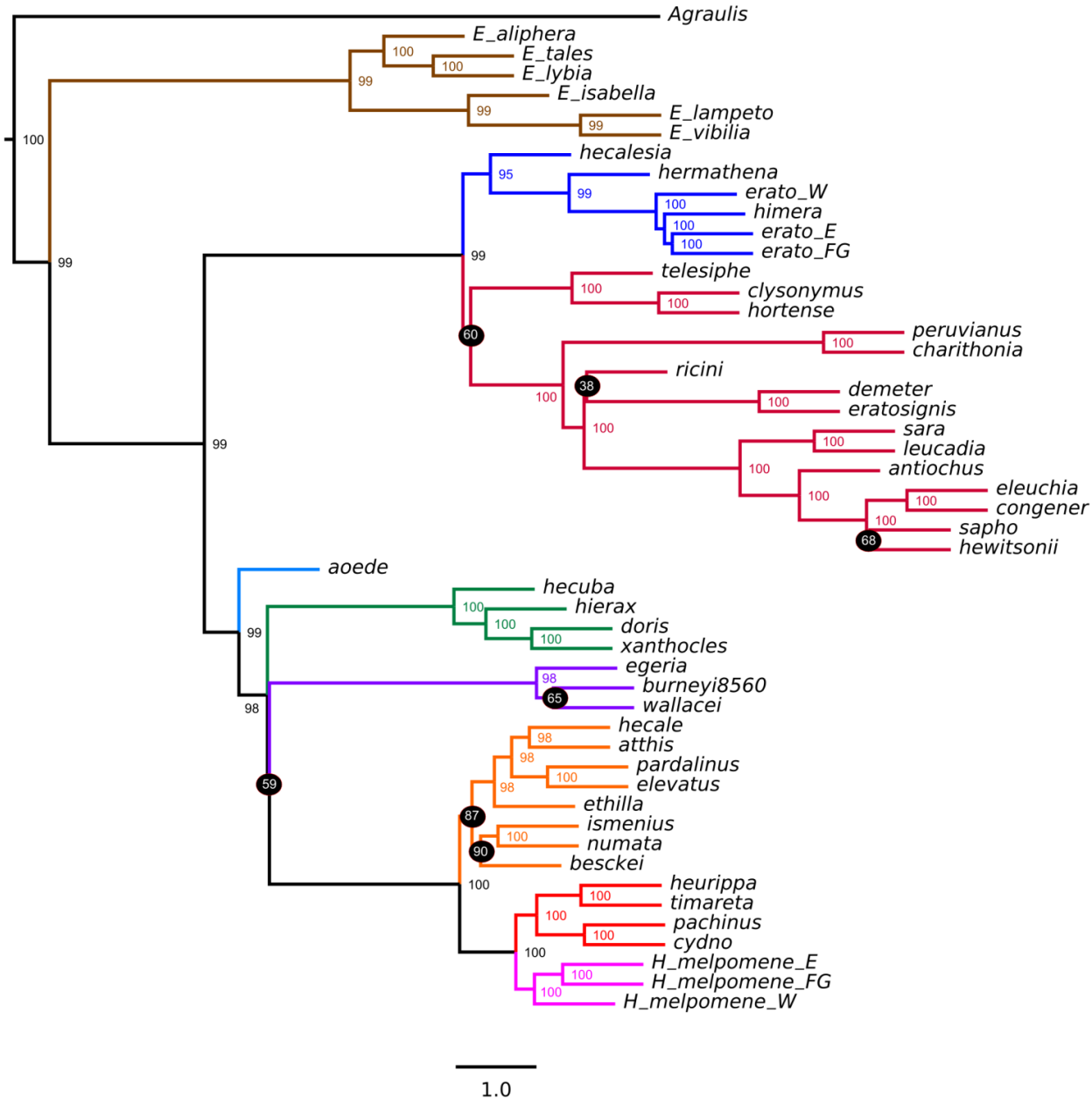


S5 Figure. Autosomal and Z-linked genes are almost equally conflicted with the species tree.

A smoothed distribution of the normalised triplet distance between the gene trees and the MP-EST species tree.



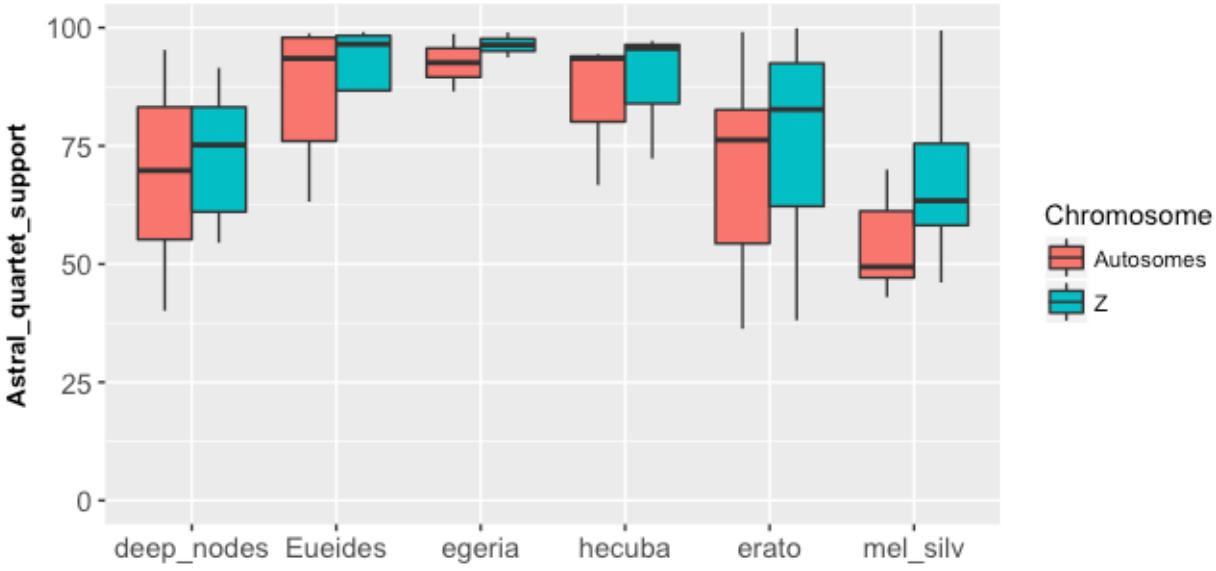
S6 Figure. Resolved and supported ASTRAL-III phylogeny for the Z (sex-linked) chromosome.



S7 Figure. Incomplete lineage sorting at the autosomal loci. A multispecies coalescent tree estimated from the 6848 autosomal CDS gene trees under the MPEST pseudolikelihood model shows lack of resolution at several nodes. Branch lengths in coalescent units, terminal branch lengths arbitrarily set to 1.0. Bootstrap support values indicated.

S8 Figure. Whole-mitochondrial sequences "resolve" the radiation of *Heliconius*, although they conflict with autosomal and sex-linked signals. A Maximum Likelihood tree (RAxML) with

bootstrap support values. Colours indicate major clades *sensu* Brown (1981). Scale bar in substitutions per site.



S9 Figure. Gene tree incongruence is the highest in the *H. melpomene*/silvaniform clade. Incongruence is higher at the autosomes than the Z sex chromosome. Mean proportion of gene tree quartets supporting all nodes as calculated in ASTRAL is plotted for all groups of *Heliconius*; *Eueides*; and the deep nodes that link them. Higher support values indicate less incongruence between individual gene trees.

BIBLIOGRAPHY

Abbott, R. et al., 2013. Hybridization and speciation. *Journal of evolutionary biology*, 26(2), pp.229–46. Available at: http://doi.wiley.com/10.1111/j.1420-9101.2012.02599.x [Accessed July 15, 2014].

Abbott, R.J., Barton, N.H. & Good, J.M., 2016. Genomics of hybridization and its evolutionary consequences. *Molecular Ecology*, 25(11), pp.2325–2332. Available at: http://doi.wiley.com/10.1111/mec.13685 [Accessed January 12, 2018].

Anisimova, M. & Gascuel, O., 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic biology*, 55(4), pp.539–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16785212 [Accessed October 31, 2014].

van der Auwera, G.A. et al., 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, 43(11.10), pp.1–33.

Bastide, P. et al., 2017. Phylogenetic Comparative Methods on Phylogenetic Networks with Reticulations. *bioRxiv*, p.194050. Available at: https://www.biorxiv.org/content/early/2017/09/28/194050?rss=1&utm_source=dlvr.it&utm_medium=twitter [Accessed January 9, 2018].

Beltran, M. et al., 2007. Do pollen feeding, pupal-mating and larval gregariousness have a single origin in {Heliconius} butterflies? {Inferences} from multilocus {DNA} sequence data. *Biological Journal of the Linnean Society*, 92, pp.221–239. Available at: //000249562600003.

Beltrán, M. et al., 2002. Phylogenetic discordance at the species boundary: comparative gene genealogies among rapidly radiating Heliconius butterflies. *Molecular biology and evolution*, 19(12), pp.2176–90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12446809.

Brand, C.L. et al., 2013. A selective sweep across species boundaries in Drosophila. *Molecular biology and evolution*, 30(9), pp.2177–86. Available at: http://mbe.oxfordjournals.org/content/30/9/2177 [Accessed March 14, 2015].

Brawand, D. et al., 2014. The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, 513(7518), pp.375–381. Available at: http://dx.doi.org/10.1038/nature13726 [Accessed September 3, 2014].

Briscoe, A.D. et al., 2013. Female Behaviour Drives Expression and Evolution of Gustatory Receptors in Butterflies. *PLoS Genetics*, 9(7), p.e1003620.

Brower, A.V.Z. & Garzón-Orduña, I.J., 2017. Missing data, clade support and "reticulation": the molecular systematics of *Heliconius* and related genera (Lepidoptera: Nymphalidae) re-examined. *Cladistics*. Available at: http://doi.wiley.com/10.1111/cla.12198 [Accessed May 12, 2017].

Brown, K. & Benson, W.W., 1975. West Colombian biogeography. Notes on Heliconius hecalesia and H. sapho (Nymphalidae). *Journal of the Lepidopterists' Society*, 29, pp.199–212.

Brown, K.S. & Benson, W.W., 1977. Evolution in modern {Amazonian} non-forest islands: {Heliconius} hermathena. *Biotropica*, 9, pp.95–117.

Brudno, M. et al., 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome research*, 13(4), pp.721–31. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=430158&tool=pmcentrez&rendertype=abstract [Accessed March 20, 2014].

Bull, V. et al., 2006. Polyphyly and gene flow between non-sibling Heliconius species. *BMC biology*, 4, p.11. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16630334.

Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, 25(15), pp.1972–3. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2712344&tool=pmcentrez&rendertype=abstract [Accessed October 25, 2014].

Chang, C.C. et al., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), p.7. Available at: https://academic.oup.com/gigascience/article-lookup/doi/10.1186/s13742-015-0047-8 [Accessed November 8, 2017].

Clarkson, C.S. et al., 2014. Adaptive introgression between Anopheles sibling species eliminates a major genomic island but not reproductive isolation. *Nature communications*, 5, p.4248. Available at:

http://www.nature.com/ncomms/2014/140625/ncomms5248/full/ncomms5248.html [Accessed March 20, 2015].

Criscuolo, A. & Gribaldo, S., 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC evolutionary biology*, 10(1), p.210. Available at: http://www.biomedcentral.com/1471-2148/10/210 [Accessed October 17, 2014].

Cui, R. et al., 2013. Phylogenomics reveals extensive reticulate evolution in Xiphophorus fishes. *Evolution; international journal of organic evolution*, 67(8), pp.2166–79. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23888843 [Accessed January 13, 2015].

Dasmahapatra, K.K. et al., 2007. Genetic analysis of a wild-caught hybrid between non-sister Heliconius butterfly species. *Biology letters*, 3(6), pp.660–3. Available at: //168.14.0.16 [Accessed June 29, 2011].

Davey, J.W., 2013. heliconius.org: Aligning Heliconius short read sequences. Available at: http://www.heliconius.org/2013/aligning-heliconius-short-read-sequences/ [Accessed May 1, 2013].

DePristo, M.A. et al., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5), pp.491–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3083463&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Durand, E.Y. et al., 2011. Testing for ancient admixture between closely related populations. *Molecular biology and evolution*, 28(8), pp.2239–52. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3144383&tool=pmcentrez&rendertype=abstract.

Edwards, S. V et al., 2016. phylogenetics continuum. , 113(29).

Feliner, G.N. et al., 2017. Is homoploid hybrid speciation that rare? An empiricist's view. *Heredity 2017 118:6*, 118(6), pp.513–516. Available at: http://www.nature.com/doifinder/10.1038/hdy.2017.7 [Accessed January 12, 2018].

Fennell, T., 2010. Picard Tools. Available at: broadinstitute.github.io/picard.

Fontaine, M.C. et al., 2014. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, p.science.1258524-. Available at: http://www.sciencemag.org/content/early/2014/11/25/science.1258524.full [Accessed November 27, 2014].

Gaunt, T.R., Rodríguez, S. & Day, I.N., 2007. Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool "CubeX." *BMC Bioinformatics*, 8(1), p.428. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-428 [Accessed January 10, 2018].

Gilbert, L.E., 2003. Adaptive novelty through introgression in Heliconius wing patterns: Evidence for shared genetic "tool box" from synthetic hybrid zones and a theory of diversification. In C. L. Boggs, W. B. Watt, & P. R. Ehrlich, eds. *Ecology and Evolution Taking Flight: Butterflies as Model Systems*. Chicago: Univ. of Chicago Press.

Hahn, M.W. & Nakhleh, L., 2016. Irrational exuberance for resolved species trees. *Evolution; international journal of organic evolution*, 70(1), pp.7–17. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26639662 [Accessed September 14, 2016].

Heled, J. & Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data. *Molecular biology and evolution*, 27(3), pp.570–80. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2822290&tool=pmcentrez&rendertype=abstract [Accessed June 13, 2011].

Heliconius Genome Consortium, 2012. Islands of divergence underlie adaptive radiation in a butterfly genome. *Nature*, 487(7405), pp.94–98. Available at: http://dx.doi.org/10.1038/nature11041 [Accessed October 25, 2012].

Hurst, G.D.D. & Jiggins, F.M., 2005. Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proceedings. Biological sciences / The Royal Society*, 272(1572), pp.1525–34. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-26244452955&partnerID=tZOtx3y1 [Accessed

July 10, 2014].

Jarvis, E.D., 2016. Perspectives from the Avian Phylogenomics Project: Questions that Can Be Answered with Sequencing All Genomes of a Vertebrate Class. *Annual Review of Animal Biosciences*, 4(1), pp.45–59. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26884102 [Accessed September 10, 2018].

Jarvis, E.D. et al., 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215), pp.1320–1331. Available at: http://www.sciencemag.org/content/346/6215/1320.abstract [Accessed December 11, 2014].

Jay, P. et al., 2018. Supergene evolution triggered by the introgression of a chromosomal inversion. *Current Biology*.

Jombart, T. & Ahmed, I., 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics (Oxford, England)*, 27(21), pp.3070–1. Available at: http://bioinformatics.oxfordjournals.org/content/27/21/3070 [Accessed September 10, 2014].

Kang, J.H. et al., 2013. Comprehensive phylogenetic analysis of all species of swordtails and platies (Pisces: Genus Xiphophorus) uncovers a hybrid origin of a swordtail fish, Xiphophorus monticolus, and demonstrates that the sexually selected sword originated in the ancestral li. *BMC evolutionary biology*, 13(1), p.25. Available at: http://www.biomedcentral.com/1471-2148/13/25 [Accessed March 21, 2015].

Kozak, K.M. et al., 2015. Multilocus Species Trees Show the Recent Adaptive Radiation of the Mimetic Heliconius Butterflies. *Systematic biology*, p.syv007-. Available at: http://sysbio.oxfordjournals.org/content/early/2015/01/28/sysbio.syv007.abstract?keytype=ref&ijkey=ymZnJKtsRrI2feJ [Accessed February 4, 2015].

Kronforst, M.R. et al., 2006. Multilocus analyses of admixture and introgression among hybridizing {Heliconius} butterflies. *Evolution*, 60, pp.1254–1268. Available at: //000238969900013.

Kronforst, M.R. & Papa, R., 2015. The functional basis of wing patterning in Heliconius butterflies: the molecules behind mimicry. *Genetics*, 200(1), pp.1–19. Available at: http://www.genetics.org/content/200/1/1 [Accessed October 3, 2016].

Kronforst, M.R.R. et al., 2013. Hybridization Reveals the Evolving Genomic Architecture of Speciation. *Cell reports*, 5(3), pp.666–77. Available at: http://www.cell.com/article/S2211124713005652/fulltext [Accessed January 16, 2014].

Lamichhaney, S. et al., 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, 518(7539), pp.371–375. Available at: http://www.nature.com/articles/nature14181 [Accessed September 3, 2018].

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), pp.2078–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), pp.1754–60. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234&tool=pmcentrez&rendertype=abstract [Accessed July 17, 2011].

Li, L., Stoeckert, C.J. & Roos, D.S., 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, 13(9), pp.2178–2189. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12952885 [Accessed January 10, 2018].

Liu, L., Yu, L. & Edwards, S. V, 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC evolutionary biology*, 10(1), p.302. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=4&SID=T1641fpcK73If782fJ8&page=1&doc=1 [Accessed March 23, 2013].

Long, C. & Kubatko, L., 2017. The effect of gene flow on coalescent-based species-tree inference. Available at: http://arxiv.org/abs/1710.03806 [Accessed January 9, 2018].

Lunter, G. & Goodson, M., 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research*, 21(6), pp.936–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20980556 [Accessed July 18, 2011].

Mallet, J. et al., 2007. Natural hybridization in heliconiine butterflies: the species boundary as a continuum.

Bmc Evolutionary Biology, 7, p. Available at: //000244937000001.

Martin, A. et al., 2012. Diversification of complex butterfly wing patterns by repeated regulatory evolution of a {Wnt} ligand. *Proceedings of the National Academy of Sciences of the United States of America*, 109(31), pp.12632–12637. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3411988&tool=pmcentrez&rendertype=abstract [Accessed October 9, 2013].

Martin, S.H. et al., 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome research*, 23(11), pp.1817–28. Available at: http://genome.cshlp.org/content/early/2013/09/17/gr.159426.113 [Accessed October 1, 2013].

Martin, S.H., 2017. Genomics general scripts. Available at: https://github.com/simonhmartin.

Martin, S.H. et al., 2018. Recombination rate variation shapes barriers to introgression across butterfly genomes. *bioRxiv*, (63).

Martin, S.H., Davey, J.W. & Jiggins, C.D., 2015. Evaluating the Use of ABBA-BABA Statistics to Locate Introgressed Loci. *Molecular biology and evolution*, p.msu269-. Available at: http://mbe.oxfordjournals.org/content/early/2014/10/14/molbev.msu269.full [Accessed November 27, 2014].

Mavarez, J. et al., 2006. Speciation by hybridization in {Heliconius} butterflies. *Nature*, 441, pp.868–871. Available at: //000238254100041.

Mavárez, J. & Linares, M., 2008. Homoploid hybrid speciation in animals. *Molecular Ecology*, 17(19), pp.4181–4185. Available at: http://doi.wiley.com/10.1111/j.1365-294X.2008.03898.x [Accessed November 8, 2012].

Mazo-Vargas, A. et al., 2017. Macroevolutionary shifts of WntA function potentiate butterfly wing-pattern diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 114(40), pp.10701–10706. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28923954 [Accessed January 13, 2018].

McKenna, A. et al., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9), pp.1297–303. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2928508&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Mérot, C. et al., 2013. Genetic differentiation without mimicry shift in a pair of hybridizing Heliconius species (Lepidoptera: Nymphalidae). *Biological Journal of the Linnean Society*, 109(4), pp.830–847. Available at: http://doi.wiley.com/10.1111/bij.12091 [Accessed November 14, 2013].

Meyer, M. et al., 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science (New York, N.Y.)*, 338(6104), pp.222–6. Available at: http://www.sciencemag.org/content/338/6104/222 [Accessed February 13, 2015].

Miles, A. et al., 2017. Genetic diversity of the African malaria vector Anopheles gambiae. *Nature*, 552(7683), p.96. Available at: http://www.nature.com/doifinder/10.1038/nature24995 [Accessed January 15, 2018].

Mirarab, S. et al., 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics (Oxford, England)*, 30(17), pp.i541-8. Available at: http://bioinformatics.oxfordjournals.org/content/30/17/i541.full [Accessed November 14, 2014].

Mirarab, S. & Warnow, T., 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12), pp.i44–i52. Available at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv234 [Accessed January 10, 2018].

Nadeau, N.J. et al., 2013. Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Molecular ecology*, 22(3), pp.814–26. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22924870 [Accessed September 17, 2013].

Nadeau, N.J. et al., 2012. Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1587), pp.343–53. Available at: http://rstb.royalsocietypublishing.org/content/367/1587/343.full [Accessed May 27, 2014].

Nadeau, N.J. et al., 2014. Population genomics of parallel hybrid zones in the mimetic butterflies, H.

melpomene and H. erato. *Genome research*, 24(8), pp.1316–33. Available at: http://genome.cshlp.org/content/24/8/1316.full#T3 [Accessed September 12, 2014].

Nadeau, N.J. et al., 2016. The gene cortex controls mimicry and crypsis in butterflies and moths. *Nature*, 534(7605), pp.106–10. Available at: http://dx.doi.org/10.1038/nature17961 [Accessed October 3, 2016].

Papa, R. et al., 2008. Highly conserved gene order and numerous novel repetitive elements in genomic regions linked to wing pattern variation in Heliconius butterflies. *BMC genomics*, 9(1), p.345. Available at: http://www.biomedcentral.com/1471-2164/9/345 [Accessed May 28, 2014].

Pease, J.B. & Hahn, M.W., 2015. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. *Systematic Biology*, 64(4), pp.651–662. Available at: https://academic.oup.com/sysbio/article-lookup/doi/10.1093/sysbio/syv023 [Accessed January 12, 2018].

Peter, B.M., 2016. Admixture, Population Structure, and F-Statistics. *Genetics*, 202(4). Available at: http://www.genetics.org/content/202/4/1485 [Accessed April 6, 2017].

Pickrell, J.K. & Pritchard, J.K., 2012. Inference of population splits and mixtures from genome-wide allele frequency data. H. Tang, ed. *PLoS genetics*, 8(11), p.e1002967. Available at: http://dx.plos.org/10.1371/journal.pgen.1002967 [Accessed July 11, 2014].

Price, M.N., Dehal, P.S. & Arkin, A.P., 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. A. F. Y. Poon, ed. *PloS one*, 5(3), p.e9490. Available at: http://dx.plos.org/10.1371/journal.pone.0009490 [Accessed July 10, 2014].

Purcell, S., 2009. PLINK. Available at: http://pngu.mgh.harvard.edu/purcell/plink/.

Reddy, S. et al., 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. *Systematic Biology*, 66(5), pp.857–879.

Reich, D. et al., 2009. Reconstructing Indian population history. *Nature*, 461(7263), pp.489–94. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19779445 [Accessed February 14, 2015].

Renaut, S. et al., 2014. Genomics of homoploid hybrid speciation: diversity and transcriptional activity of long terminal repeat retrotransposons in hybrid sunflowers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1648), pp.20130345–20130345. Available at: http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2013.0345 [Accessed January 12, 2018].

Robinson, D.F. & Foulds, L.R., 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1–2), pp.131–147. Available at: http://dx.doi.org/10.1016/0025-5564(81)90043-2 [Accessed February 27, 2013].

Rosser, N. et al., 2015. Extensive range overlap between Heliconiine sister species. *BMC Evolutionary Biology. In press.*

Rosser, N. et al., 2012. Testing historical explanations for gradients in species richness in heliconiine butterflies of tropical America. *Biological Journal of the Linnean Society*, 105(3), pp.479–497. Available at: http://doi.wiley.com/10.1111/j.1095-8312.2011.01814.x [Accessed March 20, 2015].

Roure, B., Baurain, D. & Philippe, H., 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Molecular biology and evolution*, 30(1), pp.197–214. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22930702 [Accessed November 8, 2013].

Salazar, C. et al., 2010. Genetic evidence for hybrid trait speciation in heliconius butterflies. B. Walsh, ed. *PLoS genetics*, 6(4), p.e1000930. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20442862 [Accessed January 12, 2018].

Salichos, L. & Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497(7449), pp.327–31. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23657258 [Accessed July 9, 2014].

Salichos, L., Stamatakis, A. & Rokas, A., 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Molecular biology and evolution*, 31(5), pp.1261–71. Available at: http://mbe.oxfordjournals.org/content/early/2014/02/07/molbev.msu061 [Accessed November 20, 2014].

Sankararaman, S. et al., 2016. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Current biology : CB*, 26(9), pp.1241–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27032491 [Accessed September 10, 2018].

Sankararaman, S. et al., 2014. The genomic landscape of Neanderthal ancestry in present-day humans.

Nature, 507(7492), pp.354–7. Available at: http://dx.doi.org/10.1038/nature12961 [Accessed July 11, 2014].

Sayyari, E. & Mirarab, S., 2016. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Molecular Biology and Evolution*, 33(7), pp.1654–1668. Available at: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw079 [Accessed January 10, 2018].

Schumer, M., Rosenthal, G. & Andolfatto, P., 2018. What do we mean when we talk about hybrid speciation? *Heredity*.

Schumer, M., Rosenthal, G.G. & Andolfatto, P., 2014. How common is homoploid hybrid speciation? *Evolution; international journal of organic evolution*, 68(6), pp.1553–60. Available at: http://doi.wiley.com/10.1111/evo.12399 [Accessed December 24, 2014].

Scornavacca, C. & Galtier, N., 2016. Incomplete Lineage Sorting in Mammalian Phylogenomics. *Systematic Biology*, 66(1), p.syw082. Available at: https://academic.oup.com/sysbio/article-lookup/doi/10.1093/sysbio/syw082 [Accessed October 19, 2017].

Shen, X.-X. et al., 2016. Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data. *G3: Genes|Genomes|Genetics*, 6(12), pp.3927–3939. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27672114 [Accessed September 3, 2018].

Sheppard, P.M. et al., 1985. Genetics and the Evolution of Muellerian Mimicry in Heliconius Butterflies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 308(1137), pp.433–610. Available at: http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.1985.0066 [Accessed January 13, 2013].

Shimodaira, H. & Hasegawa, M., 1989. Letter to the Editor Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *DNA Sequence*, pp.1114–1116.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, 30(9), pp.1312–3. Available at: http://bioinformatics.oxfordjournals.org/content/early/2014/01/21/bioinformatics.btu033.abstract [Accessed July 15, 2014].

Stryjewski, K.F. & Sorenson, M.D., 2017. Mosaic genome evolution in a recent and rapid avian radiation. *Nature Ecology & Evolution*, 1(12), pp.1912–1922. Available at: http://www.nature.com/articles/s41559-017-0364-7 [Accessed March 21, 2018].

Supple, M.A. et al., 2013. Genomic architecture of adaptive color pattern divergence and convergence in Heliconius butterflies. *Genome research*, 23(8), pp.1248–57. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23674305 [Accessed November 18, 2013].

Swofford, R., 2002. PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods).

Tamura, K. et al., 2012. Estimating divergence times in large molecular phylogenies. *Proceedings of the National Academy of Sciences of the United States of America*, 109(47), pp.19333–8. Available at: http://www.pnas.org/content/109/47/19333.full [Accessed July 15, 2014].

Tamura, K. et al., 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution*, 30(12), pp.2725–9. Available at: http://mbe.oxfordjournals.org/content/early/2013/10/16/molbev.mst197 [Accessed July 9, 2014].

Wahlberg, N. et al., 2009. Nymphalid butterflies diversify following near demise at the {Cretaceous}/{Tertiary} boundary. *Proceedings of the Royal Society B-Biological Sciences*, XX, p.XXX.

Wallbank, R.W.R. et al., 2016. Evolutionary novelty in a butterfly wing pattern through enhancer shuffling. *PLoS Biol*, 14(1), p.e1002353. Available at: http://dx.doi.org/10.1371/journal.pbio.1002353 [Accessed January 22, 2016].

Wen, D. et al., 2018. Inferring Phylogenetic Networks Using PhyloNet. *Systematic Biology*, 0(May), pp.1–6. Available at: https://academic.oup.com/sysbio/advance-article/doi/10.1093/sysbio/syy015/4921127 [Accessed March 21, 2018].

Wiens, J.J. & Morrill, M.C., 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Systematic biology*, 60(5), pp.719–31. Available at: http://sysbio.oxfordjournals.org/content/early/2011/03/27/sysbio.syr025 [Accessed November 7, 2013].

Xia, Q. et al., 2004. A draft sequence for the genome of the domesticated silkworm (Bombyx mori). *Science*

(New York, N.Y.), 306(5703), pp.1937–40. Available at: http://www.sciencemag.org/content/306/5703/1937.abstract [Accessed October 31, 2014].

Zhan, S. et al., 2014. The genetics of monarch butterfly migration and warning colouration. *Nature*, 514(7522), pp.317–321. Available at: http://www.nature.com/nature/journal/v514/n7522/fig_tab/nature13812_SF1.html [Accessed October 1, 2014].

Zhan, S. et al., 2011. The monarch butterfly genome yields insights into long-distance migration. *Cell*, 147(5), pp.1171–85. Available at: http://www.cell.com/article/S0092867411012682/fulltext [Accessed August 28, 2014].

Zhang, C., Sayyari, E. & Mirarab, S., 2017. ASTRAL-III: Increased Scalability and Impacts of Contracting Low Support Branches. In Springer, Cham, pp. 53–75. Available at: http://link.springer.com/10.1007/978-3-319-67979-2_4 [Accessed January 10, 2018].

Zhang, L., Mazo-Vargas, A. & Reed, R.D., 2017. Single master regulatory gene coordinates the evolution and development of butterfly color and iridescence. *Proceedings of the National Academy of Sciences of the United States of America*, 114(40), pp.10707–10712. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28923944 [Accessed January 13, 2018].

Zhang, W., Kunte, K. & Kronforst, M.R., 2013. Genome-wide characterization of adaptation and speciation in tiger swallowtail butterflies using de novo transcriptome assemblies. *Genome biology and evolution*, 5(6), pp.1233–45. Available at: http://gbe.oxfordjournals.org/content/5/6/1233.full [Accessed January 22, 2014].