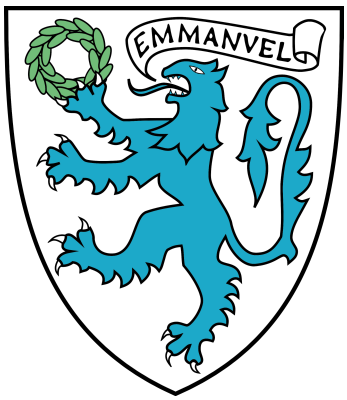
Macroevolution and phylogenomics in the adaptive radiation of Heliconiini butterflies

Krzysztof Marek Kozak



Emmanuel College, University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy in Zoology.

June 2015

DEDICATION

This work is dedicated to two people: my mother, who gave me so much,
and my sister Asia, who I wish could read it.

Doktorat dedykuję dwojgu bliskich: mojej Mamie,
oraz mojej Siostrze Asi.

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit of 60,000 words for the Degree Committee in Biology.

Krzysztof M. Kozak

26 June 2015

SUMMARY

The recognition of ecological speciation and hybridisation as key components of speciation has led to a major shift in evolutionary biology over the last decade. The mimetic *Heliconius* butterflies of the Neotropics have served as a prominent example of both, although the vast majority of studies have focused exclusively on very recent divergences and on colour pattern adaptation, neglecting deeper timescales and patterns across the rich diversity of the adaptive radiation. The relative importance across adaptive radiations of allopatry, changing ecological pressures, adaptive morphology and introgression promoted by natural and sexual selection remains unknown.

I combine phylogenetics, genomics and comparative approaches to elucidate the patterns and identify the key drivers of diversification in the continental-scale radiation of *Heliconius* and nine related genera. I present the first comprehensive, multilocus and time-calibrated phylogeny of the group and find that shifts in diversification rate cannot be unequivocally attributed to a single environmental factor. The potential role of coevolution with the obligatory host plants *Passiflora* is examined with the aid of a new phylogeny of the passion vines. Evidence is found for diffuse coevolution, as the diet of most Heliconiini is not predicted by their phylogeny and varies at short timescales.

Although passion vine butterflies are the leading example of speciation by hybridisation, this process has been described in only one subgenus. I utilise whole exome data to examine the morphologically suggestive case of a putative hybrid from another clade and find no evidence of introgression. The data is further used to answer long-standing questions about the origins of the most phenotypically diverse species. In the final chapter whole genome data are applied to characterise the patterns of divergence and gene flow across the entire genera *Heliconius* and *Eueides*, characterising the patterns of conflicting signal and

comparing the performance of philosophically distinct approaches to reflect the heterogeneity across the genome. I find that the phylogeny is unstable due to a combination if incomplete lineage sorting and introgression and may never be fully resolved, perhaps necessitating a network representation. Genomic admixture is a unique property of just one clade comprising a quarter of all species, and involves primarily the adaptive wing pattern loci. Surprisingly, the sex-linked Z chromosome shows a different order of speciation events.

Altogether my results show unexpectedly limited role of allopatry, geoclimatic variables and host plant adaptation in the diversification of a major insect radiation, thus confirming the importance of ecological speciation driven by selection on wing patterns. However, I also demonstrate that introgression may be less important in this group than previously thought.

ACKNOWLEDGMENTS

This work would not have been possible without the support, guidance and insights from Chris Jiggins. It was a privilege to learn the way of thinking from the best. I am especially grateful for the tolerance of my rather irregular ways. You gave me the space and trust I needed the most!

I gratefully acknowledge the funding: Harvard Herchel Smith Trust, Balfour Studentship from the Department of Zoology, the Emmanuel College External Research Scholarship and a grant from the Cambridge Philosophical Society.

It was a pleasure to work with the many members of the Butterfly Genetics Group over the years – great company and exceptional colleagues. I am especially indebted to Carolina Pardo-Díaz for an in-field introduction to *Heliconius*; Nicola Nadeau for help in the lab; John Davey, Ana Pinharanda and Richard Merrill for caring about my sanity. My time at the University was made much easier by the helpful staff at the Department of Zoology and at Emmanuel College. I owe special thanks to Jenny Barna, who facilitated my creative abuse of servers around the School of Life Sciences. Throughout my studies I have received useful advice and regular comments from my advisors, Rob Asher and Nick Mundy. John Welch and Richard Nichols provided excellent feedback on the final draft.

* * *

Everything started with my family, especially my mother, who over twenty years ago made the mistake of buying me a book with anatomical sketches of insects. Thank you for not choosing the Intro to Economics! The unconditional love of my mother, grandparents, uncle and aunt has always sustained me.

Eloïse Lebrun has been there for me with love and care through most of my endeavours. Thank you for keeping my interest in small rodents alive and making more cheese toast than I deserve.

Finally, a special thanks to the Emmanuel College MCR and the Cambridge Polish Society. I will miss the good times.

COLLABORATIONS AND PUBLICATIONS

I am grateful to the following collaborators for the high-throughput DNA sequencing data used in Chapters 2, 4 and 5: W. Owen McMillan, Kanchon Dasmahapatra, James Mallet, Laura Ferguson, Camilo Salazar, Mathieu Joron, Brian Counterman, Marcus Kronforst and Gilson Moreira.

Chapter 2 has been published as: **Kozak KM**, Wahlberg N, Neild AF, Dasmahapatra KK, Mallet J, Jiggins CD. ***Multilocus species trees show the recent adaptive radiation of the mimetic Heliconius butterflies.*** Syst. Biol. 2015 May;64(3):505-24.

The Bayesian dating analysis was further included in: Rosser N, **Kozak KM**, Phillimore AB, Mallet J. ***Extensive range overlap between heliconiine sister species: evidence for sympatric speciation in butterflies?*** BMC Evolutionary Biology. 2015. doi: 10.1186/s12862-015-0420-3

BEAST runs were executed on the XSEDE cluster. I appreciate the butterfly specimens shared by Andrew Neild, Christian Brévignon, Luis Constantino, Frank Jiggins, Mathieu Joron and the curators at Harvard Museum of Comparative Zoology, McGuire Center at University of Florida, Natural History Museum London and Naturhistoriches Museum Wien. I thank Nick Mundy for the permission to use vertebrate laboratory facilities. Niklas Wahlberg advised extensively on the analytics. Emily Hornett, David Lees and Rheza Zahiri shared helpful tips on historical DNA work. Jessica Leigh offered help with Concaterpillar. Rampal Etienne and Carlos Peña advised me on the diversification analysis. Members of the Butterfly Genetics Group, Rob Asher, Matthieu Joron, Nick Mundy, Albert Phillimore, Neil Rosser, Frank Anderson, Robb Brumfield and two anonymous reviewers provided helpful comments on the results and the manuscript.

Chapter 3: Tim Upson at the Cambridge Botanic Gardens commented on the identification and evaluation of *Passiflora* fossils. John Thompson offered useful insights on herbivore coevolution.

Chapter 4: Almost all of the work was carried out in Panama with support from a Graduate Fellowship at the Smithsonian Tropical Research Institute. Owen McMillan provided most of the Illumina data and served as an excellent mentor, whereas Megan Supple helped me with computing advice. The project was further funded by the Systematics Research Fund.

Chapter 5: Simon Martin contributed not only a plethora of scripts that enabled me to handle the genomic data more effectively, but also many hours of stimulating debate on hybridisation. Neil Rosser and Jake Morris were my rain-soaked companions while collecting rare species in Suriname, with sponsorship from the Panton Trust. Aylwyn Scally generously allowed me to use the Bonobo server. Tim Shaw offered help with the implementation of MP-EST.

FURTHER PROJECTS

Some aspects of my work are not included in this thesis, but have been published as parts of larger collaborations.

De novo assemblies of Illumina data were re-analysed as part of a study on the *Gustatory Receptor* gene family: Briscoe AD, Macias-Muñoz A, **Kozak KM**, Walters JR, Yuan F, Jamie GA, Martin SH, Dasmahapatra KK, Ferguson LC, Mallet J, Jacquin-Joly E, Jiggins CD. ***Female behaviour drives expression and evolution of Gustatory Receptors in butterflies***. PLoS Genetics. 2013. doi: 10.1371/journal.pgen.1003620

Detailed studies of the molecular clock models and population divergence in the *H. melpomene* complex were included in the following manuscript on the timing of gene flow:
Martin SH, Eriksson A, **Kozak KM**, Manica A, Jiggins CD. ***Speciation in Heliconius butterflies: Minimal contact followed by millions of generations of hybridisation.*** BioRxiv. 2015. http://dx.doi.org/10.1101/015800

Some of the ideas leading to Chapter 5 were initially tested with RAD-seq data and published as part of work on admixture in *H. melpomene* and relatives:
Nadeau NJ, Martin SH, **Kozak KM**, Salazar C, Dasmahapatra KK, Davey JW, Baxter SW, Blaxter ML, Mallet J, Jiggins CD. ***Genome-wide patterns of divergence and gene flow across a butterfly radiation.*** Mol. Ecol. 2013 Feb;22(3):814-26.

ABBREVIATIONS

ABC – Approximate Bayesian Computation
AICM – AIC analogue inferred from a Bayesian posterior
ASR – Ancestral State Reconstruction
BAC – Bacterial Artificial Chromosome
BIC – Bayesian Information Criterion
CDS – Coding DNA Sequence
CR – Conservatism Ratio (in host plant usage)
ELW – Expected Likelihood Weights test
EST – Expressed Sequence Tag
GTR – General Time-Reversible model
GWAS – Genome-Wide Association Study
HGT – Horizontal Gene Transfer
IC/ICA – Information Criterion nodal support
ILS – Incomplete Lineage Sorting
LBA – Long Branch Attraction
LRT – Likelihood Ratio Test
MCMC – Markov Chain Monte Carlo
MCS – clade of *Heliconius melpomene*/*H. cydno*/silvaniforms (e.g. *H. numata*) & cognates
MDC – Minimise Deep Coalescences (a phylogeny inference tool)
MDL – Minimum Description Length algorithm for detecting recombination
MDS – Multidimensional Scaling Ordination
ML – Maximum Likelihood
MP – Maximum Parsimony
MRC – Majority Rule Consensus
MRCA – Most Recent Common Ancestor
MSC – Multispecies Coalescent
NGS – Next Generation Sequencing
PCA – Principal Component Analysis
PCR – Polymerase Chain Reaction
RAD-seq – Restriction site-Associated DNA sequencing
SH – Shimodaira-Hasegawa test
SNP – Single Nucleotide Polymorphism
TC/TCA – Total Information Criterion nodal support
VCF – Variant Calls File
WG – Whole Genome (resequencing)

SUPPLEMENTARY DATA

Supplementary data for individual chapters (Figures, Tables, raw data files, sample lists) are available online through the FigShare depository.

Chapter 2: www.figshare.com/s/1a17c942194511e587d106ec4b8d1f61

Chapter 3: www.figshare.com/s/52b526b8194611e5853d06ec4b8d1f61

Chapter 4: www.figshare.com/s/2e82cb44194411e587d106ec4b8d1f61

Chapter 5: www.figshare.com/s/cbee6f46194411e593b206ec4b8d1f61

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION: PHYLOGENOMICS AND HYBRIDISATION IN ADAPTIVE RADIATIONS

Comparative methods underlie much of our understanding of macroevolution, but rely on robust phylogenies (Glor 2010). However, the foundation of phylogenetic inference has been recently shaken by the growing appreciation of incongruence and instability in the historical signals of speciation (Edwards 2009; Salichos and Rokas 2013; Fontaine et al. 2015). Simultaneously, hybridisation has been recognised as a prominent factor in speciation, which can serve as the source of novel adaptive variation and contribute to divergence in face of natural selection (Abbott et al. 2013). It is also one of the key processes that erode the signatures of ancestry and confound phylogeny reconstruction (Fontaine et al. 2015). The relative contribution of hybridisation to adaptive radiations remains unclear, but can be characterised and quantified with high-throughput sequencing.

Heliconius butterflies are a charismatic group in which this goal can be accomplished effectively. Found across the Neotropics, the 45 species and their relatives in nine less diverse genera of the tribe Heliconiini, are well known as the leading example of Müllerian mimicry in wing patterning, which has also contributed to the divergence of the radiation (reviewed by Jiggins 2008; Merrill et al. 2015). The evolution and genetics of the patterns have been researched extensively, including a plethora of studies utilising genome-wide data (Martin et al. 2013; Nadeau et al. 2013) and leading to the surprising conclusion that the key adaptive loci are exchanged between at least some of the species (Pardo-Díaz et al. 2012; Heliconius

Genome Consortium 2012; Wallbank et al. 2015). Thorough understanding of ecology and extensive genomic studies have made *Heliconius* the leading example of speciation by hybridisation (Jiggins et al. 2008; Schumer et al. 2014). *Heliconius* is also often invoked as one of the key examples of an adaptive radiation driven by ecology (Schluter 2000; Jiggins 2008; Nosil 2012). These two unusual aspects of speciation are synergistic, as intraspecific introgression may facilitate the spread of genotypes conferring ecological advantages (Arnold 2004; Abbott et al. 2013; Seehausen et al. 2014). However, many fundamental questions about the timing of the radiation, its response to varying ecological pressures and the extent of hybridisation in the broader macroevolutionary context remain unanswered. Instead, many aspects of heliconian biology have been understood in terms of speculative hypotheses formed before they could be tested with cladistic, genetic and genomic tools (Benson et al. 1975; Brown 1981; Gilbert 1991). In the following chapter I highlight the challenges and opportunities for studying macroevolution and introgression in the context of phylogenomics and introduce *Heliconius* as a system to investigate these problems.

Phylogenomics, hybridisation and the multispecies coalescent

The deluge of data generated by next generation sequencing technologies has posed both an exciting opportunity and a tremendous challenge for phylogenetics. The gradual development of the discipline in its modern form since the 1960s has involved a shift in preferred data sources from morphology to molecules, as well as associated changes in analytical approaches leading to the current prevalence of model-based methods (Hull 1988; Felsenstein 2004; Edwards 2009). However, the sudden arrival of unprecedented amounts of next-generation sequencing data has confronted the discipline with the need to modify its current practices dramatically (Jeffroy et al. 2006; McCormack et al. 2013; Leaché et al. 2014; Pyron et al. 2014). The central goal of phylogenetics has been to uncover the order and

timing of branching among the lineages of life and thus provide a framework for comparative studies into the evolution of biodiversity. Phylogenomics, most commonly understood as the extension of traditional phylogenetics to accommodate large amounts of markers sampled from across genomes (Jeffroy et al. 2006), provides tools to accomplish not only this basic goal, but also to investigate the more nuanced aspects of evolution, including heterogeneity of processes across the genome, emergence of gene families and regulatory networks, as well as horizontal gene transfer and hybridisation (Edwards 2009; McCormack et al. 2013). In this chapter, I review the contribution of phylogenomics to our knowledge of evolution generally and specifically in Lepidoptera. In the empirical sections of this thesis I present phylogenetic hypotheses for the tribe Heliconiini and subsets thereof based on increasing amounts of data, and focus on detecting and quantifying hybridisation across the genus *Heliconius*.

Hybridisation

The recent availability of genome resequencing technology for non-model taxa has dramatically improved our ability to detect and quantify the signatures of admixture and introgression and verify their importance in animal evolution. Gene flow has been increasingly recognised as a major force affecting adaptation and speciation in natural populations. Although less frequent in animals than in other kingdoms, genomic admixture is now suspected in about one tenth of all animal species (Abbott et al. 2013), where it occurs at many levels of divergence (Table 1.1). The importance of gene flow in speciation varies, as some gene flow complicates the inference in phylogenomics and population genomics, but does not appear to have contributed adaptive alleles and driven speciation. In contrast, the specific case of adaptive introgression also involves sharing of organellar genomes (e.g. mitochondria in the *Drosophila* clade; Bachtrog et al. 2006) or relatively small portions of the nuclear genome, which are maintained in different backgrounds due to a strong

selective advantage (e.g. the Dennis-Ray wing pattern loci exchanged between *Heliconius* butterflies diverged 4 MYA; Heliconius Genome Consortium 2012). Sometimes selection can lead to a substantial amount of genomic mosaicism, i.e. sharing of large portions of the genome, due to the advantage of combining multiple parental alleles in a novel environment (e.g. the alpine specialist butterflies *Lycaeides sp.*; Gompert et al. 2014). Genomic signatures of admixture in *Heliconius* can take many forms, from introgression of one or a few specific loci (Salazar et al. 2010; Pardo-Díaz et al. 2012) to admixture at over a third of the genome (Martin et al. 2013). Ample morphological evidence for interspecific hybridisation exists (Mallet et al. 2007), although the extent of this process across the radiation and across its genome is has not been previously characterised.

The heterogeneity of signals across the genomes, resulting from admixture, should be seen as a phenomenon of profound biological importance, rather than noise to integrate out of the data. A growing body of research suggests that admixture is not only a force breaking down barriers, but to the contrary, it often plays a creative role in speciation (Abbot et al. 2013). Nonetheless, the importance of gene flow and hybridisation for speciation must be considered carefully. Although adaptive introgression may involve loci of critical importance in ecological speciation (e.g. *Cottus* fish; Renaut 2011), and homoploid hybridisation is a plausible mode of speciation known from some animals (*Papilio appalachiensis;* Zhang et al. 2013), a genomic signature of admixture is insufficient to demonstrate the role of hybridisation in a speciation event (Schumer et al. 2014). The genomic evidence must be accompanied by an argument based on the ecology of the organism in question. Critically, the role of gene flow in speciation can be demonstrated only if the horizontally transmitted loci had to encode traits under strong selection leading to reproductive isolation of the incipient lineage (Jiggins et al. 2008; Schumer et al. 2014).

Taxon	Number of hybridising lineages	Divergence time	Data	Evidence	Questions addressed	References
Heliconius melpomene – H. timareta – H. elevatus	3 species	3.8 MYA	Sequence capture; WG resequencing	Phenotype; gene tree topology; ABBA-BABA tests; development	Role of adaptive introgression in reproductive isolation, during and post post speciation	Heliconius Genome Consortium 2012, Wallbank et al. 2015
Heliconius heurippa	2 species	0.43 MYA	Adaptive locus-linked and neutral nuclear loci	LD, gene trees	The role of Homoploid Hybrid Speciation (HHS)	Salazar et al. 2010
H. melpomene clade	5 species, multiple races	2.2 MYA	Adaptive locus-linked and neutral nuclear loci	IM, gene trees	Adaptive introgression between established species	Pardo-Díaz et al. 2012
H. melpomene clade	5 species	2.2 MYA	Sequence capture; WG resequencing	Gene tree topology; ABBA-BABA tests; genetic distance	Non-adaptive gene flow during and after speciation	Nadeau et al. 2012, 2013; Kronforst et al. 2013; Martin et al. 2013
Lycaeides butterflies	3 species	2.4 MYA	Reduced genomic complexity sequencing, restriction-fragment sequencing	PCA; Bayesian genomic cline analysis (Gompert and Buerkle 2011); GWAS	Introgression, HHS	Gompert et al. 2013, 2014; Nice et al. 2013
Papilio appalachiensis (swallow tail butterfly)	3 species	0.6 MYA	RNA-seq	Phylogenetic analysis of conserved loci; ABBA-BABA	HHS	Kunte et al. 2011, Zhang et al. 2013
Drosophila pseudoobscura – D. persimilis	2 species	0.2 MYA	WG resequencing		Gene flow between species	Kulathinal et al. 2009
D. subobscura – D. madeirensis	2 species	0.8 MYA	Sanger sequencing of autosomal and sex-linked loci	IM; gene trees	Gene flow (*RpS26* locus)	Khadem et al. 2011; Herrig et al. 2014

Taxon	Lineages	Divergence time	Data	Evidence	Questions addressed	References
D. simulans – D. mauritiana – D. sechellia	3 species	0.24 MYA	WG resequencing	Maximum Likelihood modelling	Recent gene flow between species	Garrigan et al. 2012; Brand et al. 2013
Gryllus pensilvanicus – G. firmus (cricket)	2 species		454 transcriptome sequencing, SNP capture	Cline model	Gene flow (restricted)	Andrés et al. 2013
Xiphophorus (swordtail fish)	2 hybridisations between species	2.4 MYA	RAD-seq (mapped to reference genome)	Species tree (supermatrix), genotypic distance	Gene flow and organellar introgression between species	Jones et al. 2013
Xiphophorus	7 hybridisations	2.4 MYA	RNA-seq	MSC; ABBA-BABA	Gene flow, HHS	Cui et al. 2013, Schumer et al. 2013
Lake Victoria Cichlidae (cichlid fish)	2 hybridisations between species	0.2-0.1 MYA (Genner et al. 2007)	RAD-seq	Outlier scan; gene trees at divergent loci	Interspecific gene flow	Wagner et al. 2013
Poecilia phormosa (Amazon molly fish)	2 species		WG sequencing	Bayesian cline analysis; genotypic distance	HHS	Alberici da Barbiano et al. 2013
Cottus rhenanus – C. perifretum (invasive sculpins)	2 species	2.0 MYA	Reduced representation sequencing; Expressed Sequence Tags	Population trees, PCA; Expression patterns	Homoploid hybridisation without complete isolation	Stemshorn et al. 2011; Czypionka et al. 2012; Cheng et al. 2013
Oncorhynchus mykiss – O. clarkii (trout)	2 species	<500 years	Overlapping paired-end RAD-seq	Outlier analysis	Very recent gene flow between established species	Hohenlohe et al. 2013; Hand et al. 2015
Passer italiae (the Italian sparrow)	2 species		RNA-seq	Genomic cline analysis	HHS	Hermansen et al. 2014; Trier et al. 2014
Anatinae (dabbling ducks)	5 species hybridising, 1 hybrid species	13.5 MYA (Gonzalez et al. 2009)	SNP capture	STRUCTURE, PCA	Homoploid hybridisation, gene flow between species	Kraus et al. 2012

Taxon	Lineages	Divergence time	Data	Evidence	Questions addressed	References
Mus musculus musculus – M. m. domesticus	2 subspecies		Mouse Diversity Genotyping Array (SNPs)	Cline analysis; gene trees, outlier analysis,	Gene flow, adaptive introgression between lineages without reproductive isolation	Teeter et al. 2010; Baird et al. 2012; Staubach et al. 2012
Canis	3 species	1.0 MYA	SNP capture	PCA, STRUCTURE, allele sharing trees, LD	Gene flow post speciation	von Holdt et al. 2011; Monzón et al. 2014
Homo sapiens – H. neanderthalensis	2 species	0.8 MYA	Capture, WG resequencing	PCA, STRUCTURE, ABBA-BABA, LD	1-4% gene flow, adaptive introgression	Green et al. 2010; Durand et al. 2011; Eriksson and Manica 2012; Sankararaman et al. 2014
Homo sapiens – Denisovan	2 species	0.8 MYA	WG sequencing, GBS	Haplotype networks, PCA, ABBA-BABA	1-6% Gene flow, adaptive introgression	Reich et al. 2010; Mendez et al. 2012; Meyer et al. 2012; Huerta-Sánchez et al. 2014
Meloidogyne incognita (root knot nematode)	2 species		WG sequencing (draft)	Gene trees	HHS	Lunt et al. 2014
Manacus vitellinus – M. candei (manakin birds)	2 species		Enzyme digest GBS	Outlier locus, PCA, Bayesian cline	Gene flow across the genome	Parchman et al. 2013
Anopheles gambiae (malaria mosquito)	2 divergent forms		WG resequencing; SNP genotyping	LD, tree scans	97% of the genomes exchanged between species	Weetman et al. 2012; Lee et al. 2013, 2014; Fontaine et al. 2015
Anopheles gambiae – A. coluzzii (malaria mosquito)	2 species	<100 years	WG resequencing	Outlier analysis,	Role of recent strong selection in driving introgression	Clarkson et al. 2014; Norris et al. 2015
Pseudacris clarkii – P. maculata (frogs)	2 species		Reduced Representation Library sequencing	Bayesian MSC	Mitochondrial introgression	Barrow et al. 2014

Taxon	Lineages	Divergence time	Data	Evidence	Questions addressed	References
Ciona intestinalis (sea squirt)	2 cryptic species	3.8 MYA	RNA-seq	ABC simulation of speciation models	5-8% gene flow between species	Roux et al. 2013
Sus (pigs)	5 species	4.2 MYA	WG resequencing	ABBA-BABA; MSC	Gene flow post speciation	Frantz et al. 2013
Corvus cornix – C. corone (hooded and carrion crows)	2 species		WG sequencing, RNA-seq	Outlier analysis, gene expression,	Gene flow across the genome	Poelstra et al. 2014
Darwin's finches	3 species	0.9 MYA	WG resequencing	LD, tree scans, ABBA-BABA	Introgression (*ALX1*) contributing to adaptation and reproductive isolation	Lamichhaney et al. 2015
Brown bear – polar bear	2 species	0.4 MYA	WG resequencing	ABBA-BABA	8.8% genome-wide gene flow	Cahill et al. 2015

Table 1. Instances of hybridisation between established or incipient species in animals, investigated at the genome level. WG=whole genome; RAD-seq: Restriction site Associated Digest sequencing; GBS: Genotyping-By-Sequencing (without a reference genome); IM: Isolation-with-Migration; LD: Linkage Disequilibrium measures; MSC: Multispecies Coalescent phylogenetic analysis; ABBA-BABA: the site-based quartet test of introgression of Green et al. 2010; PCA: Prinicipal Component Analysis; HHS: Homoploid Hybrid Speciation.

Data drives progress

Phylogenomic analyses driven by novel types of data have swept through the field of phylogenetics and been widely applied to address a variety of outstanding questions (McCormack et al. 2013). Initial efforts dealt primarily with evolution at deeper timescales and focused on widely shared sets of commonly expressed genes, captured as ESTs and later through the RNA-seq technology. Although many impressive efforts have been made with large matrices generated by Sanger sequencing of nuclear DNA (e.g. Regier et al. 2010), high-throughput transcriptomics has played a pivotal role in the epic efforts to resolve and time the origins of invertebrate phyla, particularly insects (Dunn et al. 2008; Rota-Stabelli et al. 2010; Misof et al. 2014), by providing orders of magnitude more data at many levels of sequence variation.

Whereas transcriptomics is only an intermediate step for phylogenomics, great advances have been occurring in the production and analysis of whole-genome data, initially driven by the biomedical interest in mammalian genomes and increasingly extended to the wider diversity of life, for example through initiatives to sequence ten thousand vertebrate (Genome 10K Community of Scientists 2009), five thousand arthropod (Robinson et al. 2011) and a thousand plant genomes (Matasci et al. 2014). The debates raging over the instabilities in the mammalian phylogeny have benefited immensely from these advances (Ranwez et al. 2007; Hallström and Janke 2010; Song et al. 2012; Gatesy and Springer 2014), and rapid progress is being made for other groups (e.g. Jarvis et al. 2014 on birds). Availability of entire genomes is especially advantageous for an accurate orthology prediction, and serves as a source of information on structural features like position of transposable elements and indels, considered as “molecular morphology” characters that arise infrequently compared to single site substitutions. However, the cost and effort involved in the production of a quality genome remain limiting factors (Gayral et al. 2013). For insects, genome-scale data at lower

taxonomic levels is available for only a few clades of biomedical interest, including *Drosophila* (Clark et al. 2007) and *Anopheles* (Fontaine et al. 2015).

More cost-effective ways of generating large datasets have been quickly developed for populations and phylogenetic genomics, focusing on subsampling of the genome ("genotyping by sequencing"). Identification of large sets of conserved orthologous regions has facilitated the development of capture-based approaches, such as Anchored Phylogenomics (Lemmon and Lemmon 2012), Ultraconserved Elements (Smith et al. 2013), or Hybrid Sequencing for plants (Weitemier et al. 2014), whereby known sequences from the closest reference genome are used as baits to physically bind homologous fragments of DNA from other taxa. This approach has been rapidly adapted mostly for vertebrate radiations (e.g. Pyron et al. 2013 on squamates), and enables powerful studies of recent radiations (e.g. Smith et al. 2013 on Neotropical birds). Arguably, the most insights have come from the reference-free sub-representation approaches, typically based on the sequencing and *de novo* assembly of fragments adjacent to a widespread enzyme cut-site, as exemplified by the ubiquitous RAD-seq technology (Eaton and Ree 2013). Albeit applied widely, rapid sequence divergence at noncoding loci makes RAD-seq suitable only for studies of recently diverged taxa, as demonstrated by the non-alignability of *Heliconius* RAD data between species diverged over 5 MYA (Nadeau et al. 2013). A cheap and simple alternative, developed especially vigorously in Lepidoptera, is to focus sequencing efforts on the organellar sequences, for example the mitogenome, which can be targeted by a combination of Sanger and next-generation amplicon sequencing (Timmermans et al. 2014). However, mitogenomes are often only good as an approximation due to their limited length and unalignability of some sections at deeper timescales (Wu et al. 2014).

Models and the philosophy of inference

Despite the promising developments, or rather because of them, phylogenomics has created its own problems, particularly the need to propose computationally-tractable models that capture the diversity of signals in the data (Jeffroy et al. 2006). Steel (2005) classifies the modelling problems in phylogenetics into three broad categories: model unidentifiability (overfitting); systematic error due to models that do not represent the evolutionary processes correctly; and sampling error due to insufficient number of either characters or taxa. The super-exponential increase in the amount of data largely eliminates the problem of insufficient character sampling and the associated limitations to model fitting as the characters vastly outnumber parameters, although the model complexity involved some approaches has kept up pace with the data (e.g. Martin et al. 2015). The problem of insufficient taxon sampling is potentially worsened by the much greater investment needed to add genomic data for more taxa. In Chapter 2 of this work I present a solution based on composite data.

The issue of inadequate modelling has become especially pressing, since parsing gargantuan amounts of data through simplistic models can lead to inflated false confidence from low estimates of error around inaccurately estimated means (Kumar et al. 2012). One example are the disagreements over the relative utility of supermatrix approaches and more complex coalescent modelling (see Chapter 2), recently crystallised in the debate over the mammalian tree (Hallström and Janke 2010; Song et al. 2012; Gatesy and Springer 2013, 2014). The representation of multiple heterogeneous regions in the genome, rather than a few loci, creates the need to accommodate the differences in history resulting from incomplete lineage sorting, ancient polymorphism, rate variability, effects of gene duplication, recombination, hybridisation and gene flow (see Chapter 5), which vastly exceeds the capacities of most commonly applied models (Edwards 2009; Cutter 2013). In essence, a bifurcating tree may be far from an ideal representation of the evolutionary process (Huson

and Bryant 2006; Hallström and Janke 2010).

Incorporating all the complicating factors into phylogenetic models remains an impossibility, although careful balancing acts have produced algorithms that incorporate one or two of the above. Based on the firm theoretical foundation of the coalescent theory, a large class of multispecies coalescent approaches tackles the issues of ancient polymorphism and incomplete lineage sorting (reviewed in Chapter 2; Knowles and Kubatko 2010 and papers therein), also increasingly attempting to incorporate the possibility of either gene duplication and loss (Boussau et al. 2013) or introgression (Yu et al. 2013, 2014), but never both (reviewed by Nakhleh 2013). However, the intensive studies of the human evolutionary past, characterised by an odd combination of sequential bottlenecks, worldwide expansion and periodic introgression (reviewed by Ermini et al. 2014), have stimulated the development of computational approaches to dissect out various features of population genetics, especially gene flow (Pickrell and Pritchard 2012). Increasingly, some contemporary workers assume that population-level processes can be extrapolated to the species level in recent radiation (Edwards 2009; Larget et al. 2010; Cutter 2013) and leave signatures even in very old clades (Song et al. 2012). Thus coalescent techniques, intended to study gene trees at the population level, are also useful for studying species complexes like *Heliconius.* Nonetheless, modelling co-variance between ILS and hybridisation at the scale of genomes appears to be limited by computation time (Yu et al. 2013) and restricted to cases of four to five taxa (Pease and Hahn 2014; Yu et al. 2014). In the final empirical chapter of this dissertation I explore various sources of data and analytical approaches to quantify the relative contribution of these processes in *Heliconius*.

Genomics of butterfly patterning

The greatest advances in lepidopteran genomics came arguably in the area of identifying and dissecting the genetic pathways that produce wing patterns, which have historically stimulated the interest in butterfly evolution, ecology and development (Merian 1705; Bates 1863; Reed and Nagy 2005; Gallant et al. 2014). Remarkably, studies of development suggest deep conservation of the lepidopteran patterning *Bauplan* and repeated use of the same genes at the scale of millions of years (Nijhout 1991; Reed et al. 2009; Martin et al. 2012; Martin and Orgogozo 2013; Quah et al. 2015), including the deep conservation of the *poikilomeusa* (*Yb*) gene as the melanin switch in both *Heliconius* (Nadeau et al. 2014) and the iconic peppered moth *Biston betularia* (van't Hof et al. 2011).

In *Heliconius*, genetic crosses (e.g. Sheppard et al. 1985) and transcriptomics (e.g. Hines et al. 2012; Pardo-Díaz & Jiggins 2014) led to the discovery of a small number of loci governing the basic elements that combine to generate hundreds of patterns, five of which (Table 1.2) have so far been localised in the genome by QTL mapping (Kapan et al. 2006; Baxter et al. 2009; Counterman et al. 2010; Chamberlain et al. 2011; Joron et al. 2011; Papa et al. 2013) and GWAS (Heliconius Genome Consortium 2012; Supple et al. 2013; Nadeau et al. 2014; Wallbank et al. 2015). Even though the genetic architecture of the strongly predator-selected traits has so far proven highly similar in the divergent *H. erato* and *H. melpomene* lineages (Hines et al. 2011; Martin et al. 2012; Supple et al. 2013), it remains to be seen whether the same holds for the rest of the genus and its close relatives, many of which – such as the Old World Acraeini and the sympatric co-mimics Ithomiini (Elias et al. 2008; Bernaud 2015) – bear similar markings. In at least one case the genomic organisation of the colour loci appears to have been highly modified, as *H. numata* evolved a single supergene *P* in a large inversion (Joron et al. 2011), which results in a unique diversity of the characteristic tiger forms (Fig. 2.1).

The wing patterns of *Heliconius* are among the most striking examples of adaptive introgression and even prior to the use of genomic data it became clear that the evolution of the involved genes is largely independent from the history of the rest of the genome (Salazar et al. 2010, Hines et al. 2011, Pardo-Díaz et al. 2012). The genomic perspective has been indispensable in quantifying the amount of introgression and narrowing down the involved loci (Wallbank et al. 2015). Throughout this dissertation I use phylogenomics to study a wider range of colour pattern loci than before across a greater number of *Heliconius* species.

Locus						
H. melpomene	***H. erato***	**Other**	**Phenotype**	**Scaffold**	**Genes**	**bp**
B	*D*	*Br/G*[1], *P*[2]	Red on HW and FW, ventral brown patterns	HE670865	*optix, kinesin,* Dennis/Ray enhancers	602276
Yb/Sb/N	*Cr*	*P*	Yellow/white on HW and FW	HE667780	*poikilomeusa*	1333114
Ac	*Sd*	*P*	FW melanism and band shape	HE668478 HE669520	*wntA*	521029 205163
Ro	*Ro*	*P*	Shape of FW band	HE671554	*radial spoke head 3*	334452
K	*K*	*P*	White/yellow switch	HE671246 HE670889	*aristaless 1*, *aristaless 2*	93691 79976

Table 1.2. The five major wing pattern and colour loci of *Heliconius*. The key loci are named differently in the *H. erato* and *H. melpomene* species. Main genes likely to be implicated in pattern formation listed. Length of scaffolds and number of sliding windows reported. [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.

THESIS OUTLINE

This thesis consists of four empirical chapters intended as individual manuscripts. The overarching goal of the projects is to understand the patterns and test the hypotheses regarding potential drivers of diversity in *Heliconius* and more broadly in the tribe Heliconiini. Thus I return the macroevolutionary perspective on heliconian evolution, blending phylogenomics, molecular evolution and comparative methods.

1. The first study attempts to reconstruct the phylogeny of the tribe Heliconiini based on a combination of Sanger sequencing from modern and historical specimens with *de novo* assembly of Illumina data. Special attention is paid to the ongoing debates on the relative merits of various computational approaches in phylogenetics. Finally, I investigate the changes in the diversification rate across the tribe and their potential links to the dynamic environmental change in the Neotropics.

2. In the second chapter I synthesise our knowledge of associations between Heliconiini and their host plant group Passifloraceae. Although the importance of passion vines in heliconian evolution is widely acknowledged, their phylogenetic patterns of diversification have never been compared to determine the extent to which the two clades have influenced one another's evolution. Using a novel, comprehensive phylogeny of Passifloraceae I quantify the extent of codivergence between the two clades and investigate a range of hypotheses regarding their coevolution.

3. The third chapter shifts the emphasis to population-level processes, recent divergences and hybridisation. I use genome-wide sequencing data to investigate the possibility that *Heliconius hermathena* originated by homoploid hybrid speciation, which could potentially

be the first case outside the intensively studied *melpomene*/silvaniform clade. The dataset is further used to address the long-standing debate about the origin and dispersal of the most variable and widespread species in the genus, *Heliconius erato*.

4. In the final study I revisit the challenge of estimating a phylogeny for the fast-evolving, hybridising genus *Heliconius*. Whole genome data are generated for the majority of species in the radiation. I combine several approaches to identify the sources of instability in the phylogeny and disentangle hybridisation from other processes, while comparing functionally different parts of the genome.

MULTILOCUS SPECIES TREES SHOW THE RECENT ADAPTIVE RADIATION OF THE MIMETIC *HELICONIUS* BUTTERFLIES

Visual mimicry provides an excellent system in which to study the origins of biodiversity, as the targets of selection are clearly identifiable and the role of natural selection in promoting adaptation and ultimately speciation can be directly observed (Bates 1863; Benson 1972; Mallet and Barton 1989; Sherratt 2008; Pfennig and Editor 2012). Studies of mimetic assemblages have been instrumental in explaining many biological phenomena, but testing hypotheses regarding the evolution of mimicry depends heavily on our knowledge of systematic relationships between the participating taxa, particularly where mimics are closely related (Ceccarelli and Crozier 2007; Wright 2011; Penney et al. 2012). Unfortunately, both strong selection on adaptive loci, which may facilitate adaptive introgression, and rapid radiation leading to incomplete lineage sorting, are likely to be common in many systems, but are especially prevalent in mimetic butterflies (Savage and Mullen 2009; Kubatko and Meng 2010; Kunte et al. 2011; Zhang et al. 2013b). These processes can significantly interfere with the estimation of phylogenies (Maddison and Knowles 2006; Linnen and Farrell 2008; Edwards 2009; Anderson et al. 2012).

A large body of recent work has been devoted to the issue of incongruence between the species tree (the true speciation history) and the gene trees evolving within (Anderson et al. 2012; Cutter 2013). The traditional approach of concatenating the total genetic evidence into a supermatrix to obtain a global estimate of the predominant phylogenetic signal and

hidden support (Gatesy and Baker 2005), without consideration for the heterogeneity of individual partitions, has been to some extent superseded by multispecies coalescent (MSC) techniques (reviewed in: Edwards 2009; Knowles and Kubatko 2010; Anderson et al. 2012; Cutter 2013; Leaché et al. 2013). The majority of the new MSC algorithms are intended to model at least some of the sources of heterogeneity between different markers, most frequently focusing on the problem of incomplete lineage sorting (e.g. Maddison and Knowles 2006; Heled and Drummond 2010), sometimes addressing hybridization (e.g. Gerard et al. 2011; Yu et al. 2011), and in at least one case modelling discordance without specifying its potential source (Larget et al. 2010). Although the supermatrix approach remains popular and serves as an effective approximation of the species diversification history in most cases, its ability to properly assess the degree of statistical support for phylogenies has been brought into question and contrasted with the potential of the MSC techniques to assess confidence more realistically (Edwards 2009; Knowles and Kubatko 2010). Heliconiini are an especially interesting subject for a systematic study, where the purported robustness of MSC tools to gene flow and other sources of incongruence can be tested with real data.

I estimate the phylogeny and investigate the link between the dynamics of speciation and macroevolutionary factors in the Neotropical butterfly tribe Heliconiini (Nymphalidae: Heliconiinae). Heliconiini are arguably the most thoroughly-researched example of microevolution in the Neotropics, the most biologically diverse region of the world (Hoorn et al. 2010). They comprise the genus *Heliconius* and nine smaller genera, providing a spectacular example of a radiation where speciation is promoted by divergence in mimicry of aposematic wing patterns (Jiggins et al. 2001b; Arias et al. 2008; Merrill et al. 2012). Heliconiini, and especially *Heliconius*, form Müllerian mimicry rings of distantly related toxic species, in which mimetic species share the cost of educating avian predators by

evolving similar wing patterns. Thus, Heliconiini are an excellent system for the study of convergence from both genomic and organismal perspectives (e.g. Duenez-Guzman et al. 2009; Hines et al. 2011; Bybee et al. 2012; Heliconius Genome Consortium 2012; Pardo-Díaz et al. 2012; Jones et al. 2013; Martin et al. 2013; Supple et al. 2013; Arias et al. 2014). In addition to studies of wing pattern evolution, there has been a recent proliferation of comparative molecular and genomic studies of other traits including vision (Pohl et al. 2009), chemosensation (Briscoe et al. 2013) and cyanogenesis (Chauhan et al. 2013). A robust and stable molecular phylogeny is therefore especially desirable for this clade.

Heliconius, the most species-rich genus of Heliconiini, has the key features of an adaptive radiation (Schluter 2000; Glor 2010). All *Heliconius* are found in broadly similar Neotropical habitats (Brown 1981) and most of the species appear to originate on the eastern slope of the Central and North Andes (Rosser et al. 2012). *Heliconius* is far more diverse than the other nine genera of the tribe, and individual species show high disparity in their dietary adaptations (Benson et al. 1975; Merrill et al. 2013) and wing pattern diversity (Brown 1981; Jiggins 2008a). Numerous field and lab experiments have demonstrated the adaptive value of wing patterns in protection from predators (Langham 2004). Importantly, divergence in wing pattern also leads to both pre- and post-mating reproductive isolation and hence contributes to speciation (Jiggins et al. 2001b; Merrill et al. 2012). There is also evidence that ecological divergence in mating behaviour, preference for microhabitat, and host plant use permit sympatric co-occurrence of closely related species (Gilbert 1991; Mallet and Gilbert 1995; Jiggins et al. 1997b). There is also evidence for putative key innovations, most notably pollen feeding, which enables unusually long lifespans and underlies much of the unusual ecology of the group (Cardoso and Gilbert 2013). An expansion of olfactory receptor gene families (Briscoe et al. 2013) seems likely to play a role in specific host and food plant preferences (Benson et al. 1975; Boggs et al. 1981; Merrill et al. 2013). In summary, *Heliconius* have

undergone a burst of diversification associated with adaptive changes, some of which directly cause reproductive isolation. Nonetheless, one characteristic of adaptive radiations that has not yet been demonstrated in *Heliconius* is a temporal burst of species diversification (Glor 2010; Moen and Morlon 2014).

An unusual feature of *Heliconius* is the prevalence and importance of gene flow and hybridization, leading to a controversy over the validity of traditional species concepts in the clade (Beltrán et al. 2002; Mallet et al. 2007). At least 26% of all species of Heliconiini occasionally produce interspecific hybrids in the wild (Mallet et al. 2007), and at least one species, *H. heurippa*, has resulted from homoploid hybrid speciation from parental forms diverged millions of years ago (Mavárez et al. 2006; Jiggins et al. 2008; Salazar et al. 2010). Genome sequencing has shown that mimetic diversity in the *H. melpomene* and silvaniform clades is also explained by adaptive introgression of genes regulating the aposematic wing patterns (Heliconius Genome Consortium 2012; Pardo-Díaz et al. 2012). In addition, neutral gene flow seems to be widespread, influencing as much as 40% of the genome between *H. melpomene* and *H. cydno* (Martin et al. 2013).

Heliconiini systematics has a long history, starting with early morphological work (Emsley 1963; Brown 1981; Penz 1999), through allozymes (Turner et al. 1979) and a combination of morphological and ribosomal DNA-restriction data (Lee et al. 1992), to studies based on the sequences of mitochondrial and nuclear markers (Brower 1994b; Brower and Egan 1997; Beltrán et al. 2002, 2007; Cuthill and Charleston 2012; Massardo et al. 2014). The most comprehensive study to date by Beltrán et al. (2007) attempted to address some of the difficulties by incorporating many taxa (38 of 46 *Heliconius*, 59 of 77 Heliconiini), considering two individuals of most species, and sequencing two mitochondrial (*CoI/II*, *16S*) and four nuclear markers (*EF1α*, *Wg*, *Ap*, *Dpp*). However, this dataset is still potentially inadequate to address the challenges posed by Heliconiini systematics, as three of the loci

(*16S*, *Ap*, *Dpp*) were only sequenced for 12 representative species. Among the other three markers, 65% of the variable sites resided in the fast-evolving mitochondrial partition *CoI/II*, raising the possibility that the inferred relations are largely driven by the historical signal of the matriline. The relationships between *Heliconius*, *Eueides* and the other eight genera were not resolved with good support, as might be expected if most of the phylogenetically informative variation comes from a fast-evolving partition. Importantly, the individual data partitions were analysed in concatenation. Despite these shortcomings, Beltrán and colleagues confirmed that the morphological and behavioural characteristics of the major clades do not correspond directly to their evolutionary history and that the traditionally recognized genera *Laparus* and *Neruda* are most likely nested within the crown genus *Heliconius*.

The importance of Müllerian mimicry as a driver of individual speciation events has been well-established (Mallet et al. 1998; Jiggins 2008a) whereas the macroevolutionary processes governing the evolution of the group have been largely neglected in empirical studies, with the exception of two recent studies on density-dependent speciation (Etienne et al. 2012) and biogeography of speciation (Rosser et al. 2012). Precise understanding of the evolution of the mimicry rings, as well as associated processes such as hybridization, requires knowledge of the relative timing of the divergence events and motivated the most widely-cited study of the molecular clock in Arthropoda (Brower 1994b). Mallet et al. (2007) created a chronogram from a partially unresolved tree, using a relaxed clock procedure and the *CoI/II* alignment. Importantly, the first dated phylogeny of Heliconiini is not calibrated to an absolute standard, making it impossible to make inferences about the relation of the diversification process and the contemporaneous geological and climatic events. A recent comparative study suggests that the majority of *Heliconius* lineages originated in the north-eastern Andes and spread to other parts of the continent (Rosser et al. 2012). Recent reviews (Hoorn et al. 2010; Rull 2011; Turchetto-Zolett et al. 2013), based on the cumulative results of

up to 200 systematic studies, suggest that most South American tropical clades have experienced periods of significantly elevated net diversification rate in response to Andean orogenesis, alterations in the hydrology and sediment dynamics of the present-day Amazon Basin, as well as local and global climatic changes. These processes can result in allopatry of incipient lineages, or in creation of new ecological niches and change to the species-level carrying capacity of the environment leading to ecological speciation (Brumfield and Edwards 2007; Hoorn et al. 2010). In particular, we can hypothesize that the diversification rate of Heliconiini increased during the periods of especially rapid Andean uplift around 23, 12 and 4.5 Ma (Gregory-Wodzicki 2000; Solomon et al. 2008; Hoorn et al. 2010).

Aims of the study

Here I aim to resolve the species tree of Heliconiini radiation and generate a dataset including nearly all of the currently recognised species in the tribe, sampling intraspecific diversity across the range of many species, combining whole genome sequencing with Sanger sequencing. I apply a wide range of phylogenetic methods to reconstruct the species tree, including supermatrix, coalescent and network approaches, which allow me to assess the strength of the underlying signal of speciation. The power of my combined approach is harnessed to elucidate the importance of marker heterogeneity for the final assessment of systematic relationships, while realistically estimating the support values for my chosen topology. I date the time of individual divergence events with sufficient precision, permitting an analysis of diversification dynamics. I thus present a comprehensive study of macroevolutionary dynamics in a mimetic system that has been studied intensively at the microevolutionary level.

MATERIALS AND METHODS

Taxon sampling

I sampled 180 individuals, including 71 of the 77 species in all genera of Heliconiini and 11 outgroup species. The specimens came primarily from the collection at the University of Cambridge, with additional specimens shared by museums and private collectors (Online Appendix 1). I included five outgroup species from the sister tribe Acraeini (Wahlberg et al. 2009) and three from the related genus *Cethosia*. The diverse analyses used in this paper require different sampling designs and the demands of all the techniques cannot be easily accommodated in a single dataset. For example, the network analysis based on nucleotide distance produced much better supported and resolved trees when the 95% or more incomplete data from historical specimens were not used, whereas the various multispecies coalescent techniques required the use of at least two individuals per species and had to be based only on taxa with intraspecific sampling (Fulton and Strobeck 2009; Heled and Drummond 2010). Thus I distinguish four datasets. The complete data matrix includes all the data. The core dataset excludes 14 individuals represented solely by short DNA fragments from historical specimens. The single-individual dataset includes both modern and historical specimens, but with only the single best-sequenced individual per taxon. Finally, the *BEAST matrix contains only the 17 species of *Eueides* and *Heliconius* with extensive sampling of multiple representatives of each species.

DNA sequencing

I used 20 nuclear and two mitochondrial loci as markers (Supplementary Tables 1 and 2; Online Appendix 1). The selection includes the three classic molecular markers for Lepidoptera (*CoI/II, EF1α*, *Wg*), two markers proposed by Beltrán et al. (2007) (*16S*, *Dpp*), eight new universal markers proposed by Wahlberg and Wheat (2008) (*ArgK*, *Cad*, *Cmdh*, *Ddc*, *Idh*, *Gapdh*, *Rps2*, *Rps5*) and nine highly variable loci identified by Salazar et al. (2010)

(*Aact*, *Cat*, *GlyRS*, *Hcl*, *Hsp40*, *Lm*, *Tada3*, *Trh*, *Vas*). Additional *Heliconius*-specific primers were designed for *Cmdh*, *Gapdh* and *Idh*. Details of the primers and PCR cycles are listed in the Supplementary Table 1. For most species, sequences of the three basic markers for multiple individuals were already published (Brower 1997; Beltrán et al. 2007; Wahlberg et al. 2009; Salazar et al. 2010), and data for 26 individuals came exclusively from GenBank (Online Appendix 1).

I generated Sanger sequences for 103 specimens (Online Appendix 1). DNA was isolated from approximately 50 µg of thorax tissue using the DNeasy Blood & Tissue kit (Qiagen, Manchester, UK). PCR was carried out in a total volume of 20 µl, containing 1x Qiagen Taq buffer (Manchester, UK), 2.5 mM $MgCl_2$, 0.5 µM of each primer, 0.2 mM dNTPs, 1 unit bovine serum albumin, 0.5 unit Qiagen Taq-Polymerase and 1 µl of the DNA extract. The following program was executed on a G-Storm cycler (Somerton, UK): denaturation 5 minutes at 94°C; 35 cycles of 30 seconds at 94°C, 30 seconds at the annealing temperature and 90 seconds at 72°C; final extension for 10 minutes at 72°C. The results were visualized by electrophoresis in 1.5% agarose gel stained with 1% ethidium bromide. PCR products were cleaned using the ExoSAP-IT system (USB, Cleveland, Ohio): 60 minutes at 37°C; 20 minutes at 80°C. I used gel purification with the Nucleo Spin Extract II kit (Macherey-Nagel, Düren, Germany) as needed. Sanger sequencing reaction was carried out with the BigDye Terminator v. 3.1 (AB, Foster City, California: 2 minutes at 94°C; 25 cycles of 10 seconds at 94°C, five seconds at 50°C and four minutes at 60°C. The products were sequenced with the ABI 3730xl DNA Analyzer at the Sequencing Facility, Department of Biochemistry, University of Cambridge. I manually inspected the traces in CodonCode v. 4.0.4 using PHRED for quality assessment (CodonCode Corporation 2012).

At the time of my Sanger sequencing effort, whole genome data generated for other studies became available from 57 individuals in 27 common species (Heliconius Genome

Consortium 2012; Briscoe et al. 2013; Supple et al. 2013; Dasmahapatra and Mallet, unpublished) (Online Appendix 1). 100 base pair reads were generated using the Illumina Genome Analyzer II platform with insert size of 300-400 bp. I performed *de novo* assembly of the short reads in the program Abyss v. 1.3 (Simpson et al. 2009). Based on previous studies (Salzberg et al. 2012; Briscoe et al. 2013) and preliminary results (S. Baxter, *pers. comm.*), I chose k-mer length of 31, minimum number of pairs n=5 and minimum mean coverage c=2 as optimal settings. The 20 nuclear markers were mined from the assemblies by megaBLAST (Camacho et al. 2009) in Geneious v. 5.5.1 (Biomatters Ltd 2012) using reference sequences from the *Heliconius melpomene* genome. The quality of the recovered sequences was assessed by alignment to previously generated amplicon sequences of the same loci from the same individuals.

Mitochondrial sequences could not be recovered by *de novo* methods, presumably because the large number of reads from the highly abundant mitochondrial DNA contained a large enough number of sequencing errors to interfere with the assembly. I reconstructed whole mitochondrial genomes of 27 species by mapping to the *Heliconius melpomene* reference (The *Heliconius* Genome Consortium, 2012), using the default settings in the Genomics Workbench v. 5.5.1 (CLCBio 2012). This data were analysed separately from the 21 locus mixed nuclear-mitochondrial alignment.

DNA sequencing: historical specimens

Short fragments of *CoI/II* and *EF1α* were sequenced from historical specimens up to 150 years of age, obtained from museum and private collections (Online Appendix 1) and processed in a vertebrate genetics laboratory to reduce the risk of contamination. Instruments and surfaces were cleaned with 5% bleach and irradiated with UV for 30 minutes prior to use. One to two legs were washed in water, immersed in liquid nitrogen in a test tube for 30

seconds and ground up, followed by an extraction into 20 µl of buffer using the QIAmp DNA Micro Kit (Qiagen, Manchester, UK). I treated every fifth extraction and every fifth PCR as a negative control with no tissue or DNA extract. PCRs were carried out in a 20 µl volume using 1 unit of Platinum HiFi Taq Polymerase (Invitrogen, London, UK) and 1x buffer, 2.5 mM $MgCl_2$, 0.5 µM of each primer, 0.2 mM dNTPs, 1 unit bovine serum albumin, sterilized DNAse-free water and 1-5 µl of the DNA extract depending on concentration. To accommodate shearing of DNA with time, I designed and applied PCR primers spanning short fragments of 200-300 bp (Online Appendix 2). I carried out amplification, product clean up and sequencing as above, partially accounting for possible cross-contamination by blasting the results against GenBank.

Alignment and gene tree estimation

Alignments for each locus were generated in CodonCode v. 3 to account for inverted and complemented sequences, and improved using MUSCLE v. 3.8 (Edgar 2004). I visualized the alignments of the coding loci (all except the mitochondrial *16S* and the *tRNA-Leu* fragment in the *CoI/II* sequence) in Mesquite v. 2.75 (Maddison and Maddison 2011) and checked translated sequences for stop codons indicating errors. The whole mitochondrial sequences were aligned to the *Acraea issoria* and *Heliconius melpomene* references (*Heliconius* Genome Consortium 2012) using the G-INS-i algorithm in MAFFT (Katoh 2002). The number of variable and parsimony informative sites was estimated for each locus in PAUP* v. 4 (Swofford 2002). Models of sequence evolution implemented in MrBayes (Ronquist and Huelsenbeck 2003) were selected in MrModelTest v. 2.3 (Posada and Crandall 1998; Nylander 2004) based on the AIC (Akaike 1974). Xia's test in DAMBE v. 4.0 (Xia and Xie 2001) demonstrated saturation in the third codon position of *CoI/II*, prompting me to treat the third codon position of the fast-evolving *CoI/II* locus as a separate partition when

estimating the gene tree. The Leucine tRNA (*tRNA-Leu*) fragment occurring in the middle of *CoI/II* displays very low variability and thus was included in one partition with the slower evolving first and second codon positions. Individual gene trees were estimated in MrBayes v. 3.1, using four runs of one chain, 10 million MCMC cycles sampled every 1000, and 2.5 million cycles discarded as burnin based on the average standard deviation of split frequencies becoming less than 0.01 and a plateau in the log likelihood values (Ronquist and Huelsenbeck 2003). The mitochondrial genes were concatenated due to their shared history, but treated as separate partitions with distinct models. All gene trees were visualized with FigTree v. 1.4 (Rambaut 2009).

Detection of Conflicting Signals

I investigated the cyto-nuclear discordance and other conflicts in the phylogenetic signal with several methods. To illustrate the global reticulate signal in the data, a NeighborNet network was built with the pairwise distances calculated under the F84 correction, the most complex model that could be fitted to the data in the program SplitsTree v. 4 (Kloepper and Huson 2008). I reduced the dataset to a single best-sequenced individual per species to exclude the reticulations resulting from the expected recombination within species. Next, the topological disparity among individual loci was illustrated using Multi-Dimensional Scaling of pairwise Robinson-Foulds distance (Robinson and Foulds 1981) between the gene trees, as estimated by TreeSetViz v. 1.0 (Hillis et al. 2005) in Mesquite. Calculation of the RF required trimming the trees to the minimal set of 54 shared taxa from 27 species, using the R package APE (Paradis et al. 2004; R Development Core Team 2008). Finally, I investigated if topologies and branch lengths of the individual loci are consistent enough to justify concatenation of the markers by means of a hierarchical likelihood ratio test in Concaterpillar v. 1.5 (Leigh et al. 2008).

Supermatrix Phylogenetics

I created a supermatrix of the 20 nuclear and 2 mitochondrial markers in Mesquite. To increase the efficiency of tree searches, optimal partitioning schemes for the **complete** and **single-individual** datasets were identified in PartitionFinder v. 1.1 (Lanfear et al. 2012), using the Bayesian Information Criterion (BIC) (Schwarz 1978) and a greedy search, followed by selection of nucleotide substitution models. Lists of optimal partitions for ML and Bayesian analyses can be found in Online Appendix 2. The Maximum Likelihood (ML) phylogeny was searched for under the GTRGAMMA model in RAxML v. 8.1 with 350 bootstrap replicates under the GTRCAT approximation (Stamatakis 2006), where the number of repetitions was determined by the bootstopping criterion (Stamatakis 2014). To explicitly test the likelihood of various hypotheses for Heliconiini phylogeny, several alternative topologies were created in Mesquite, representing previously identified groupings, as well as possible placements of the enigmatic genera *Cethosia*, *Laparus* and *Neruda* (Supplementary Table 4) (Brower 1994b; Brower and Egan 1997; Penz 1999; Penz and Peggie 2003; Beltrán et al. 2007). I then re-estimated the ML tree using each topology as a constraint. The likelihood scores of the original and *a priori* specified alternative trees were compared using the Shimodaira-Hasegawa test (Shimodaira and Hasegawa 1989) and the Expected Likelihood Weights based on 1000 bootstrap replicates (Strimmer and Rambaut 2002). A separate phylogeny was generated for the unpartitioned whole mitochondrial alignment in RAxML under the GTRGAMMA model with 1000 bootstrap replicates.

Dating the radiation

A Bayesian chronogram was estimated using the program BEAST v. 1.8 (Drummond et al. 2012). To avoid incorrect estimates of the substitution rate parameters resulting from the inclusion of multiple samples per species, this analysis was based on a pruned alignment with one individual per species. The only exception is the inclusion of three races of *H. melpomene*

and two races of *H. erato*, where deep geographical divergences are found (Quek et al. 2010). The Vagrantini sequences were not used, as BEAST can estimate the placement of the root without an outgroup (Drummond et al. 2006). Thus the analysis included 77 taxa and the optimal partitioning scheme was re-estimated appropriately (Table 2.1). I linked the topology, but modelled an uncorrelated lognormal clock and the substitution rate separately for each partition (Drummond et al. 2006). Substitution rates were drawn from the lognormal distribution with an overdispersed gamma prior on the mean of this distribution. The shape parameter was k=0.001 (0.005 for 16S and 0.1 for the third position of CoI/II), scale parameter theta=1 and starting value 0.001 for nuclear genes, and k=0.01 for the faster-evolving mitochondrial loci (van Velzen et al. 2013). I used a Birth-Death tree prior and empirical base frequencies to limit the computation time for the heavily parametrized model.

As no fossils of Heliconiini or closely related tribes are known, I used secondary calibration points from the dated phylogeny of Nymphalidae (Wahlberg et al. 2009; van Velzen et al. 2013). Prompted by the findings of Sauquet and colleagues (2012), who demonstrate the potential for biases when secondary dating information is used, I compared having a single calibration point at the root with using all of the eight known split times. I also considered the impact of modelling each prior divergence time as normally distributed with the mean found by Wahlberg et al. (2009) and the standard deviation matching the 95% credible intervals, or as uniformly distributed within the same bounds. For each of the four model combinations, four independent instances of the MCMC chain were run for 100 million generations each, sampling the posterior every 10000 generations and discarding 10 million generations as burnin after convergence assessment in Tracer v. 1.6 (Rambaut et al. 2014). Based on the marginal likelihood estimated under the AICM criterion (Baele et al. 2012), I chose the scheme with multiple calibrations modelled as uniformly distributed (Supplementary Table 3), although I found that the ages of most of the nodes do not differ by

more than 10% between the four schemes. While I recognize that the models should ideally be chosen based on the Path Sampling or Stepping Stone procedures (Baele et al. 2012), in practice I found these to be too computationally demanding for my relatively large dataset. The input .xml file generated in BEAUti v. 1.8 (Drummond et al. 2010) can be found in the Online Appendix 3. To ensure that the results are driven by the data and not the priors, I executed an empty prior run. The Maximum Clade Credibility tree with mean age of the nodes was generated using LogCombiner v. 1.8 and TreeAnalyser v. 1.8 (Drummond et al. 2012). Parallel to the SH and ELW tests, plausible alternatives to the disputed and poorly supported nodes in the Bayesian chronogram (Supplementary Table 5) were tested by calculating the posterior model ratios, a simple and robust alternative to the computationally expensive Bayes Factor calculations (Bergsten et al. 2013). A posterior sample of 1000 phylogenies was filtered in PAUP* using constraint trees designed in Mesquite and the model odds were calculated as the frequency of the alternative grouping divided by the frequency of the clade observed in the MCC chronogram.

Multispecies coalescent phylogenetics

To account for the heterogeneous phylogenetic signal resulting from gene flow, hybridization and incomplete lineage sorting, I applied a variety of multispecies coalescent (MSC) analyses that take as input both the raw alignment and individual gene trees. I first used the established method of Minimizing Deep Coalescences (MDC) (Maddison and Knowles 2006), taking advantage of a dynamic programming implementation in the package PhyloNet (Than et al. 2008). One hundred samples of 100 trees were drawn randomly without replacement from the distribution of Bayesian gene trees for the 21 loci, an MDC phylogeny was estimated for each sample and a 50% majority rule consensus was taken.

Bayesian Concordance Analysis (BCA) is an MSC method that attempts to reconcile the

genealogies of individual loci based on posterior distributions, regardless of the sources of conflict (Larget et al. 2010). BCA generates Concordance Factors (CFs), which show what proportion of loci contain a particular clade, and estimates the primary phylogenetic hypothesis from the best-supported clades. CFs offer a powerful alternative to traditional measures of support and can be conveniently estimated in the program BUCKy (Ané et al. 2007; Larget et al. 2010). I executed two runs of one million MCMC cycles in BUCKy based on the 21 posterior distributions of gene trees from MrBayes.

Another approach to the multispecies coalescent is to estimate the gene trees and the species tree simultaneously, explicitly modelling the sources of incongruence (Edwards et al. 2007). I applied this technique using *BEAST, a program harnessing the power of BEAST and simultaneously implementing a powerful MSC algorithm that estimates the species tree and the embedded gene trees, as well as the population sizes of the lineages (Heled & Drummond 2010). Effective calculation of the population size parameters requires a thorough multilocus sampling of each species in the analysis, which forced me to reduce the dataset to species with a minimum of three individuals from at least two distinct populations (the ***BEAST dataset**). The final alignment included 87 terminal taxa in 17 species of *Eueides* and *Heliconius*. I re-estimated the individual substitution models for each partition (Supplementary Table 2), used a constant population size coalescent tree model and implemented other priors as described above for BEAST. I carried out four independent runs of 500 million cycles each, sampling every 10000 cycles, generated a maximum clade credibility species tree and visually summarized the 21 gene trees by plotting in DensiTree (Bouckaert 2010). The Online Appendix 4 contains the .xml file for this analysis.

Species delineation

New species of Heliconiini continue to be described based on morphological, genetic or karyotypic evidence (Lamas et al. 2004; Constantino and Salazar 2010; Dasmahapatra et al. 2010; Moreira and Mielke 2010)(Lamas et al. 2004; Constantino and Salazar 2010; Dasmahapatra et al. 2010; Moreira and Mielke 2010), yet the validity of the taxonomic status of some of these new lineages has been disputed (J. Mallet and A. Brower, *pers. comm*.). To explore the potential for phylogenetic determination of the number of species in the tribe Heliconiini, species were delimited with the novel Bayesian Poisson Tree Process (bPTP) algorithm (Zhang et al. 2013a). Using the ML tree with multiple individuals per species, two runs of 500,000 generations were executed with the first 10% discarded as burnin based on the convergence of log likelihood values. The results were often poorly supported and included clusters and splits not compatible with the extensive knowledge of the biology of this group (Supplementary Figure 7), prompting me instead to use the accepted taxonomy of Lamas (2004) in my study of diversification dynamics in this clade.

Changes in the Diversification Rate

The formal analyses of diversification dynamics were based on the output of the Bayesian supermatrix analysis in BEAST. Initially I investigated the changes in the number of lineages by generating semi-log Lineage Through Time (LTT) plots for Heliconiini and *Heliconius* in the R package paleotree (Bapst 2012) based on a posterior sample of 1000 chronograms, plotting the 95% credible interval of the divergence time around the median curve. I also carried out preliminary analyses in the R package DDD (Etienne and Haegeman 2012), using an MCC tree from an analysis with a single root calibration. Comparisons of all possible models, including diversity-dependence and changes in rate regimes, suggested that the diversification rate of Heliconiini increased around 12 Mya, shortly before the appearance of *Heliconius*, which in turn started to diversify faster around 4.5 Mya.

To achieve a more nuanced understanding of the diversification dynamic in this clade, I chose the reversible-jump Markov Chain Monte Carlo (rjMCMC) approach implemented in the R package BAMM (Rabosky 2014; Rabosky et al. 2014a), which models both the variation in rates between lineages and the change of rates in time. BAMM is distinct from previously published approaches, as it also treats the location of diversification regimes on the tree as a variable to model (Rabosky 2014). In effect, the algorithm does not give false confidence in single, point estimates of rate shift times. To account for the uncertainty surrounding the influence of dating priors and the divergence times, I carried out analyses with four chronograms: the MCC phylogeny, the two trees corresponding to the upper and lower boundary of the 95% credible interval, as well as the MCC tree from an analysis with a single, normally distributed calibration at the root. All the phylogenies were truncated to 1 Ma before the present, to account for the uncertainty in the estimates of recent speciation (protracted speciation) and extinction rates (pull of the present; Etienne and Haegemann 2012). Appropriate speciation and extinction prior values were chosen for each tree in the BAMMtools suite (Rabosky et al. 2014). I experimented with three alternative values of the Poisson process rate prior suitable for relatively small trees (1.0, 2.0, 4.0) and discovered that the results are very similar, as evidenced by small differences in Bayes Factors.

Following the failure of automated species delimitation, incomplete sampling of the taxa was accounted for by coding in the sampling proportion of each genus and each major *Heliconius* clade according to the taxonomy of Lamas (2004), with addition of new species of *Philaethria* (Constantino and Salazar 2010) and *Neruda* (Moreira and Mielke 2010), and excluding *H. tristero* (Mérot et al. 2013) (species sampling fraction=0.904). A test with the most extreme of the plausible taxonomies (not recognizing *H. heurippa*, *Eueides emsleyi*, the undescribed *Agraulis* or the three recently proposed *Philaethria* species; sampling fraction=0.956) demonstrated that the results are robust to the assumptions on the proportion

of missing taxa. I assumed that rate changes may occur in increments of 0.5 Mya. For each of the four trees, two independent runs of four Metropolis-Coupled MCMC chains were executed for 2,200,000 generations and the first 200,000 generations were discarded as burnin. Convergence was assessed with the R package coda (Plummer et al. 2006) by plotting the log likelihood of all chains and calculating the Effective Sample Size for log likelihood and the number of rate shifts. BAMMtools were used to analyse the posterior set of 2000 models, compare the frequently observed rate regimes with Bayes Factors, identify the branches where rate shifts occurred, calculate the mean speciation, extinction and diversification rates for major clades, and plot the change of these parameters in time.

RESULTS

Sampling and DNA sequencing

I successfully combined three approaches to sequencing, which resulted in the high taxonomic coverage and intraspecific sampling necessary for the MSC methods, and a sufficient sampling of loci for each individual (Online Appendix 1). Most of the dataset consists of Sanger sequences from 108 individuals, 26 of which were already included in GenBank. I also obtained two classic lepidopteran markers *CoI/II* and *EF1α* from respectively 13 and 11 out of 14 historical specimens, thus adding eight species of Heliconiini that have not been sequenced previously.

I capitalized on the availability of Illumina data by generating *de novo* assembly contigs for an additional 57 individuals. The N50 of the Abyss assemblies ranged from 552 to 1921 bp (average 1206) and all the nuclear markers were successfully recovered by megaBLAST from every assembly. Whole mitochondrial sequences of the same individuals were recovered by read mapping, with about a 400 bp stretch of the hypervariable control region (*Heliconius*

Genome Consortium 2012) incomplete in some sequences. I obtained a depth of coverage over 100x and high confidence in the base calls due to the high copy number of mtDNA in the tissue. Finally, I included the sequences extracted from the *Heliconius melpomene* genome, as well as the previously published mitochondrial sequence of *Acraea issoria* (Heliconius Genome Consortium 2012).

The sequence data for 20 nuclear and 2 mitochondrial genes encompass 70 out of 76 (92%) of the officially recognized species of Heliconiini, including 44 out of 46 species from the focal genus *Heliconius* (Lamas et al. 2004; Beltrán et al. 2007; Mallet et al. 2007; Constantino and Salazar 2010; Moreira and Mielke 2010; Mérot et al. 2013). Although the taxonomic validity of some species is contested, I found that the diversification analysis is robust to altering the number of missing species. Some recognized taxa diverged very recently, as shown by the BEAST chronogram (Fig. 1a) in case of the *Philaethria diatonica*/*P. neildi*/*P. ostara* complex and the *Heliconius heurippa*/*H. tristero* pair, the latter of which was recently reclassified as a race of *H. timareta* (Mérot et al. 2013). However, the exact relationship between genetic differentiation and taxonomic species identity in the highly variable, mimetic Heliconiini remains unclear (e.g. Mérot et al. 2013, Nadeau et al. 2013). Importantly, 36 species are represented by multiple individuals, usually from distant populations, allowing for more accurate estimation of sequence evolution rates, and detection of species paraphyly. The number of individuals represented by each marker ranges from 40% for *Hcl* to 98% for *CoI/II*, and only four specimens are represented exclusively by mitochondrial DNA (Online Appendix 1).

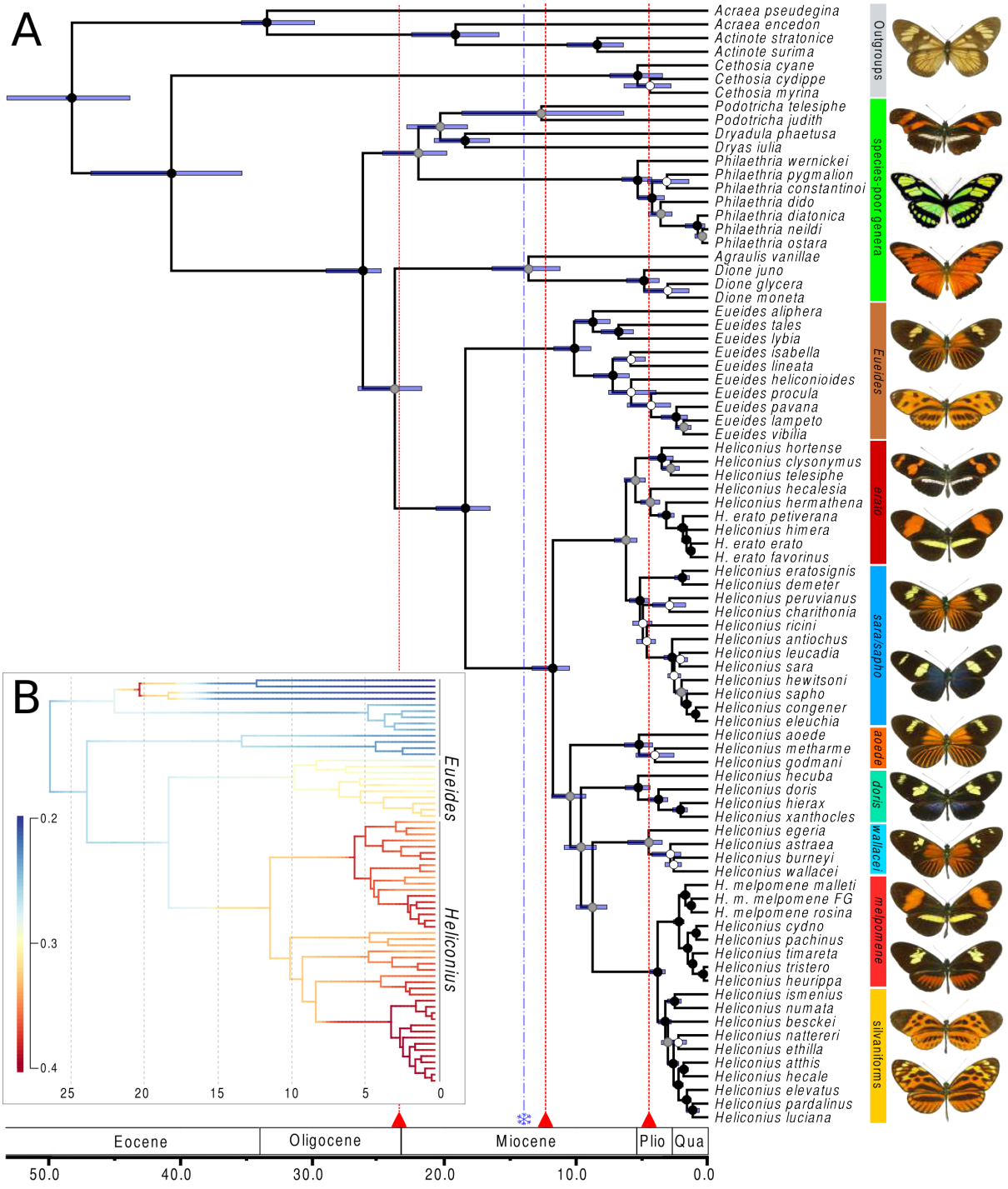

A
Acraea pseudegina
Acraea encedon
Actinote stratonice
Actinote surima
Cethosia cyane
Cethosia cydippe
Cethosia myrina
Podotricha telesiphe
Podotricha judith
Dryadula phaetusa
Dryas iulia
Philaethria wernickei
Philaethria pygmalion
Philaethria constantinoi
Philaethria dido
Philaethria diatonica
Philaethria neildi
Philaethria ostara
Agraulis vanillae
Dione juno
Dione glycera
Dione moneta
Eueides aliphera
Eueides tales
Eueides lybia
Eueides isabella
Eueides lineata
Eueides heliconioides
Eueides procula
Eueides pavana
Eueides lampeto
Eueides vibilia
Heliconius hortense
Heliconius clysonymus
Heliconius telesiphe
Heliconius hecalesia
Heliconius hermathena
H. erato petiverana
Heliconius himera
H. erato erato
H. erato favorinus
Heliconius eratosignis
Heliconius demeter
Heliconius peruvianus
Heliconius charithonia
Heliconius ricini
Heliconius antiochus
Heliconius leucadia
Heliconius sara
Heliconius hewitsoni
Heliconius sapho
Heliconius congener
Heliconius eleuchia
Heliconius aoede
Heliconius metharme
Heliconius godmani
Heliconius hecuba
Heliconius doris
Heliconius hierax
Heliconius xanthocles
Heliconius egeria
Heliconius astraea
Heliconius burneyi
Heliconius wallacei
H. melpomene malleti
H. m. melpomene FG
H. melpomene rosina
Heliconius cydno
Heliconius pachinus
Heliconius timareta
Heliconius tristero
Heliconius heurippa
Heliconius ismenius
Heliconius numata
Heliconius besckei
Heliconius nattereri
Heliconius ethilla
Heliconius atthis
Heliconius hecale
Heliconius elevatus
Heliconius pardalinus
Heliconius luciana
Outgroups
species-poor genera
Eueides
erato
sara/sapho
aoede
doris
wallacei
melpomene
silvaniforms
B
Eueides
Heliconius
0.2
0.3
0.4
25
20
15
10
5
0
Eocene
Oligocene
Miocene
Plio
Qua
50.0
40.0
30.0
20.0
10.0
0.0

Figure 2.1. (a) Bayesian phylogeny of 71 out of 77 butterflies in the tribe Heliconiini with outgroups, estimated using 20 nuclear and two mitochondrial markers with an uncorrelated molecular clock method (BEAST). The age of the root is calibrated based on the results of Wahlberg et al. (2009) and the bars signify the 95% credible intervals around the mean node ages. Scale axis in Ma. Deep splits are shown within the well-studied *Heliconius erato* and *H. melpomene*. Red triangles: mean estimates of intensive Andean orogeny (Hoorn et al. 2010). Blue flake: the end of the Miocene climatic optimum. Heliconiini exhibit complex patterns of divergence and convergence in aposematic wing patterns, top to bottom: *Actinote latior* (outgroup), *Podotricha telesiphe telesiphe*, *Philaethria dido chocoensis*, *Dione juno*, *Eueides tales michaeli*, *E. lampeto lampeto*, *Heliconius telesiphe telesiphe*, *H. erato favorinus*, *H. demeter ucayalensis*, *H. sara sara*, *H. aoede cupidineus*, *H. doris* (blue morph), *H. burneyi jamesi*, *H. melpomene amaryllis*, *H. timareta contigua*, *H. numata lyrcaeus*, *H. pardalinus dilatus*. (b) The mean phylorate plot from the BAMM analysis. Net diversification rate is averaged across 2000 models fitted to the MCC chronogram. Colours from blue to red indicate the range of diversification rates from 0.2 to 0.4 new lineages per lineage per million years. X axis in Millions of years ago. Photos © C. Jiggins, M. Joron and L. Constantino.

New DNA sequences were deposited in GenBank under Accession Numbers KP072800-KP074896 and KP113715-KP114069 (Online Appendix 1). Sequence alignments and phylogenetic trees were deposited in TreeBase (Accession Number 15531).

Pervasive conflict between the loci

The nuclear markers span 11 out of 21 chromosomes (Sup. Table 2; Heliconius Genome Consortium, 2012) and have both autosomal and sex chromosome Z-linked inheritance. I examined the conflict between individual markers in the entire tribe and in the genus *Heliconius* alone, using both gene tree summary methods and approaches utilizing the raw sequence alignments. The maximum likelihood analysis of the core matrix in Concaterpillar rejected concatenation of any of the loci due to significant differences in both topology and substitution rate of individual partitions, but the exact nature of the discordance is unclear. A Multi-Dimensional Scaling ordination of pairwise RF distances between the gene trees does not reveal clustering by chromosome and the separation between many nuclear loci appears much greater than between nuclear and mitochondrial trees (Sup. Fig. S1). Consistent with this is the fact that the whole mitochondrial phylogeny of select taxa (Fig. 2.3) shows few differences from the tree based on the mixed marker supermatrix (Fig. 2.1a), highlighting that cytonuclear discordance is not the primary source of incongruence in the dataset.

The coalescent approaches reveal the high extent of marker conflict in the *Heliconius* data. Gene tree topologies from the explicit Bayesian modelling of incomplete lineage sorting (ILS) in *BEAST are highly varied, with a particularly high degree of reticulation in the *H. melpomene*/*cydno* and the *H. hecale* (silvaniform) clades, where extensive horizontal gene flow has been observed previously (Sup. Fig. S2; Martin et al. 2013). Another Bayesian method, BUCKy, infers the species tree in the presence of marker incongruence without modelling specific reasons for the observed discordance, and calculates the Bayesian

concordance factors that illustrate the proportion of partitions in the dataset that support a particular grouping (Ané et al. 2007; Baum 2007; Larget et al. 2010). The concordance of the loci for *Heliconius* is strikingly low, although the topology is consistent with the results of other analyses (Fig. 2.1a and 2.3, Sup. Fig. S3-S7).

Further strong evidence of widespread incongruence comes from the NeighborNet network characterized by a high delta score of 0.276, which shows that the structure of the data is not entirely tree-like (Sup. Fig. S4). This can be partially attributed to the effect of missing data, yet even a fit based on the 30 species with full or nearly complete sequences produced a delta score of 0.11, proving a substantial amount of non-bifurcating signal across the tribe (Holland et al. 2002). The most noticeable reticulations are found between nodes linking genera and the major clades of *Heliconius* (Sup. Fig. S4). This is most likely due to pervasive gene flow during the diversification of the main extant lineages, as systematic error in gene tree estimation would be expected to impact nodes at different depth to a similar extent.

Topological consistency across optimality criteria

Although no trees are identical, the results from my Bayesian, Maximum Likelihood and distance-based network analyses of the supermatrix are very similar (Fig. 2.1a, Sup. Fig. S4, S5). Two basal nodes stand out as unstable. *Cethosia* is variably placed as a sister taxon to either Acraeini (MP, ML, NeighborNet) or Heliconiini (Bayesian), despite the reasonably extensive sampling of 11 loci for *C. cyane*, while the position of *Podotricha* in relation to *Dryas* and *Dryadula* varies among all analyses. Most problematic are relations among the species within *Eueides*, where the position of four out of 10 taxa cannot be resolved with good support. The poor resolution for both *Eueides* and *Podotricha* can probably be attributed to insufficient site coverage, which produces high uncertainty due to

patchily distributed missing data (Wiens and Morrill 2011; Roure et al. 2013). I re-estimated the relations of *Eueides* based solely on a core set of 11 genes with coverage for at least 7/10 species and recovered a much better supported tree (Sup. Fig. S6). I recommend studies focusing on *Eueides* use this specific phylogeny, but the exact relations of *E. procula*, *E. lineata* and *E. heliconioides* remain unclear.

Although the Bayesian maximum clade credibility tree is similar to the topology estimated by Beltrán *et al.* (2007), significantly increasing the dataset from 113 to 180 individuals and five to 21 loci allowed me to resolve many critical nodes. The major differences are in the relations of the genera other than *Heliconius*, which I infer to form a grade, with high support for *Eueides* as the sister genus of *Heliconius* (Fig. 1a). Importantly, I confirm that the enigmatic genera *Laparus* and *Neruda* are nested within *Heliconius*, as further supported by ELW and SH tests (Table 2), although *Neruda* is closer to the base of the genus than previously estimated (Fig. 1a). I find that the other so called "primitive" (Brown 1981) clade of *Heliconius* consists of two separate groups nested among other subgenera, raising questions about the apparently unequal rate of morphological evolution in the genus.

The maximum likelihood tree is similar to the Bayesian phylogeny, although it puts less confidence in the deeper nodes linking the genera and subgenera (Sup. Fig. S5), as is typically the case (Wiens and Morrill 2011). The nodes differing between the two trees are also the nodes that cannot be unequivocally confirmed by either the SH and the ELW test (Sup. Table 4). For instance, the highest posterior probability in the ELW test (0.302) is given to the tree that does not lump *Neruda* and the "primitive" *Heliconius*, consistent with the BEAST chronogram in Figure 2.1a. Thus I suggest that the Bayesian tree should be preferred as a more accurate picture of the systematic relationships, although the ML phylogeny based on the **complete** dataset is still useful to uncover multiple polyphyletic species. Notably, *H. luciana* is nested within *H. elevatus*, and *H. wallacei i*s polyphyletic with respect to

H. astraea. These results must be interpreted with caution, as the inference relies on poorly covered museum specimens and may be sensitive to long branch attraction (Wiens and Morrill 2011).

My whole mitochondrial phylogeny is largely consistent with the results of the multilocus supermatrix analysis and well-supported for 46/57 nodes (Fig. 2.3), despite the relatively limited taxonomic coverage of only 29 Heliconiini for which short-read data was available. In contrast to the multilocus dataset, mitochondrial genomes are not very useful for resolving relationships between major clades within *Heliconius*, as the positions of *Neruda*, *H. xanthocles* and *H. wallacei* clades are poorly supported. An important deviation from the predominantly nuclear multilocus phylogeny is the placement of *H. pachinus* as a sister taxon of *H. timareta*, rather than *H. cydno*. This may reflect the overall instability resulting from a high extent of reticulation in the *melpomene*/*cydno*/*timareta* assemblage (*Heliconius* Genome Consortium 2012, Martin et al. 2013, Mérot et al. 2013). Furthermore, I find a surprising positioning of *H. hermathena* within *H. erato* (Jiggins et al. 2008), which is also not supported by any of the analyses of the 21 locus matrix. The mitochondrial tree confirms previous observations of deep biogeographical splits in the widespread, highly diversified *H. erato* and its co-mimic *H. melpomene* (Brower 1994a). Within *H. melpomene* I find a well-supported distinction between races found to the west and east of the Andes, although my data place the individuals from French Guiana with the specimens from the Western clade, in contrast to a whole-genome phylogeny (Nadeau et al. 2013) (Fig. 2.1a). *Heliconius erato* shows the opposite pattern in both mitochondrial and nuclear data, whereby the Guianian races form a fully supported clade in the monophyletic group of taxa from East of the Andes.

The variety of multispecies coalescent (MSC) methods that I applied brings a new perspective to the phylogenetic signals in Heliconiini. I analysed the small dataset of the 17 best-sampled species with the Bayesian MSC algorithm implemented in *BEAST and

recovered a tree which agrees with the supermatrix analyses (Sup. Fig. S3), except for the position of *Neruda aoede*, which was placed as a sister taxon to the *H. xanthocles*/*L. doris* clade with relatively low posterior probability. Furthermore, the mean ages of nodes are very close to those proposed in a supermatrix analysis (Table 2.1), with similar 95% credible intervals. Although the species tree is largely as predicted, I observe high levels of incongruence in the underlying distribution of gene trees (Sup. Fig. S2). Differences in the depth of coalescence are clear throughout the tree and reticulation is again especially apparent in the *H. melpomene/cydno* clade. The estimated population size values are also consistent with a previous comparison based on two nuclear loci, showing a higher population size of *H. erato* (1.33×10^6 individuals) when compared to *H. melpomene* (1.02×10^6) (Flanagan et al. 2004).

Similarly, I find the phylogeny derived by the gene tree summary approach BUCKy to be entirely consistent with the Bayesian analysis of concatenated sequence, although the recovered CFs are much lower than any other measure of support applied to my data. Importantly, most of the nodes connecting the major clades in the tree have CFs below 0.5, with the notable exception of the silvaniform/*melpomene* split (Fig. 2.3). The same nodes correspond to the reticulations in the NeighborNet analysis (Sup. Fig. S4), cases of low support in the MDC tree and its disagreement with the supermatrix analysis (Sup. Fig. S7), nodes that cannot be rejected in the ML and Bayesian tests of topologies (Sup. Tables 4-5), and the uncertain nodes in the whole mitochondrial tree (Fig. 2.2).

Another summary analysis by Minimizing Deep Coalescences (MDC), although taking only a few minutes with 100 bootstrap replicates of the complete dataset, returns a very different topology from the other techniques, showing a number of unexpected and poorly supported groupings. The lumping of all non-*Heliconius* genera, and the monophyletic *Neruda*/*xanthocles*/*wallacei* clade stand out in contrast to other proposed trees (Sup. Fig. S7).

Interestingly, many of the relations that are poorly supported in the supermatrix phylogenies are also not resolved in the consensus MDC tree, showing that MDC is highly conservative with regard to the placement of taxa unstable in individual gene trees.

	Heliconius* vs *Eueides	***Heliconius* vs *Agraulis***	***Heliconius* vs *Philaethria***	***melpomene* vs *erato***	***erato* vs *hecalesia***
Mallet et al. 2007	11.0	13.0	14.5	9.5	4.0
Pohl et al. 2009	n/a	~32 (~26-38)	n/a	~17 (~12-22)	n/a
Wahlberg et al. 2009	18.5 (12.2-23.9)	26.5 (21.0-31.6)	30.0 (24.6-36.2)	n/a	n/a
Cuthill and Charleston 2012	~13 (~10.5-16.5)	n/a	n/a	~10.5 (~8.0-13.5)	~4.5 (~2.7-6.3)
BEAST	18.4 (16.5-20.6)	23.8 (21.7-26.6)	26.2 (24.8-29.0)	11.8 (10.5-13.4)	4.4 (3.7-5.1)
*BEAST	17.5 (11.6-23.5)	n/a	n/a	10.5 (6.7-14.9)	n/a

Table 2.1. Mean split ages in Ma at different levels of divergence within Heliconiini, as estimated by previous studies and by two Bayesian relaxed clock methods in the present work. 95% credible intervals and Highest Posterior Densities reported if known.

Tempo of diversification

The phylogeny estimated under a relaxed clock model in BEAST shows diversification dynamics that differ from previous estimates, with the deeper splits between the genera of Heliconiini estimated as substantially older than previously inferred with mitochondrial data, but younger than estimated with a small sample of nuclear genes (Table 2.1) (Mallet et al. 2007, Pohl et al. 2009). The most species-rich genera *Heliconius* and

Eueides separated 18.5 (95% Highest Posterior Density: 16.5-20.4) Ma and both started to diversify respectively 11.8 (10.5-13.4) Ma and 10.2 (8.9-11.7) Ma. The six major clades of *Heliconius* (corresponding to *H. erato, H. sara, H. xanthocles, H. wallacei, H. melpomene*/silvaniforms and *Neruda*) all started to diversify around 5 Ma (Fig. 1b, Sup. Fig. S8b). The extinction rate has remained relatively constant throughout the history of the tribe, while the speciation rate (and thus net diversification) increased substantially in *Eueides* and *Heliconius* (Table 2.3). The Lineage Through Time plot (Sup. Fig. S8a) and the plot of diversification rate for Heliconiini (Sup. Fig. S9) suggest rapid early emergence of *Podotricha*, *Dryadula* and *Dryas*, which later speciated at a very low rate (Fig. 2.1, Table 2.2). This is followed by a period of stasis 18-12 Ma, roughly corresponding to the mid-Miocene (Hoorn et al. 2010), and a sudden but steady increase in the rate of diversification after 11 Ma. In the case of the 45 *Heliconius*, a shorter plateau is found between 9 and 6 Ma, and the number of extant lineages rises from 5 Ma onwards (Sup. Fig. S8b). As expected, the LTT plots drop off sharply in the last million years, reflecting protracted speciation (Etienne and Haegemann 2012).

Maximum Likelihood modelling in DDD strongly supports the Birth-Death model with an increase in diversification rates 11 Ma as the best fit for Heliconiini (Akaike weight of 0.92; Sup. Table 6). However, I find that the confidence in the point estimate of shift time is largely a function of forcing the model over the data (Rabosky et al. 2014a). The Bayesian analysis with BAMM demonstrates that although there has likely been a large change in rate regimes across the phylogeny, it cannot be reduced to a single event at a strictly specified time, and this result is consistent for all four chronograms that I tested (values are reported for the MCC chronogram based on multiple uniform calibrations). Models with one or two shifts in diversification rate have the highest frequency in the posterior (0.430 and 0.281, respectively) and Bayes Factors provide weak evidence for models with one shift (BF=3.495)

or two shifts (BF=4.398) over models with no shift. Painting the mean rate per branch from 2000 individual models onto the phylogeny (Fig. 2.1b) and explicit testing of rate variation between clades (Sup. Fig. S10) show clearly that the rate changed on the branches leading to (i) the *Dryas* and *Podotricha* clade (BF>40); (ii) the *Heliconius* and *Eueides* clade (BF>10); (iii) *Heliconius*. But even though there is evidence for rate increase on the *Heliconius* branch, no particular model with a specific shift time occurs at high frequency. On the contrary, the most frequent model (f=0.41) in the 95% credible set is the one with a constant rate.

Clade	**Speciation rate**	**Extinction rate**
Heliconiini	0.267 (0.191-0.369)	0.130 (0.030-0.269)
Heliconius	0.318 (0.215-0.450)	0.122 (0.012-0.281)
non-*Heliconius*	0.236 (0.132-0.361)	0.135 (0.024-0.288)
Eueides	0.263 (0.125-0.396)	0.122 (0.012-0.275)
the eight species-poor genera	0.224 (0.111-0.356)	0.142 (0.024-0.316)

Table 2.2. Speciation and extinction rate estimates for Heliconiini butterflies calculated in the BAMM analysis of macroevolutionary rates based on a posterior sample of 1000 BEAST chronograms. Means and 90% Highest Posterior Density intervals are reported for the focal clades and for all of Heliconiini excluding the most species-rich genus *Heliconius.*

DISCUSSION

Stable topology despite marker incongruence

I approached the problem of phylogeny reconstruction in a difficult mimetic assemblage through extensive intraspecific sampling of 22 markers from nearly all species in the clade, and compared the results between multiple philosophically distinct analytical approaches. As next generation sequencing technologies become widely accessible and the average number of loci used in systematic studies increases rapidly, Multispecies Coalescent (MSC) methods gain in importance as means of detecting and accounting for incongruence in multilocus data (e.g. Lee et al. 2012; Barker et al. 2013; Smith et al. 2013). However, their relative merits and utility at different levels remain contested (Song et al. 2012; Gatesy and Springer 2013; Reid et al. 2013). The systematic relations of the tribe Heliconiini, which diverged from its extant outgroup about 47 Ma, can be effectively resolved with both MSC and supermatrix approaches, yielding highly similar topologies across a range of different sub-sampling schemes that correspond to the requirements of individual algorithms (Fig. 2.1a, 2.2 and 2.4; Sup. Fig. S3-S7). Nonetheless, consistent with the recent radiation of the group, large effective population sizes and known hybridization between many species, I observed high heterogeneity among sampled fragments of the genome that differ markedly in both topology (Sup. Fig. S2) and rates of evolution (Concaterpillar analysis). Such heterogeneity might have been expected to pose a significant challenge for the concatenation methods (Degnan and Rosenberg 2006; Edwards et al. 2007; Edwards 2009; Knowles and Kubatko 2010; Leaché and Rannala 2010). However, the only method producing an obviously different phylogeny is Minimizing Deep Coalescences (MDC; Sup. Fig. S7), which fails to resolve 34 out of 62 nodes in the tree with bootstrap support above 0.9, and is the only method to suggest monophyly of the non-*Heliconius* genera. MDC derives a species tree from point estimates of gene trees and can be expected to perform poorly with a relatively limited number of gene

trees that are not always fully resolved, leading to a complete polytomy in some of the clades (Gatesy and Springer 2013). However, the MDC result is an indicator of instability, as well-resolved and consistent gene trees should produce a good quality MDC tree.

Recent studies have proposed that likelihood-based MSC techniques should be preferred to integration over individual gene trees, due to their potential to capture synergistic effects between partitions (Leaché and Rannala 2010; Reid et al. 2014), mirroring the phenomenon of hidden support in concatenation (Gatesy and Baker 2005). My results support this observation and further show that high degrees of conflict between many partitions can be reconciled by both supermatrix and MSC approaches to extract the predominant signal of speciation. The Bayesian concordance analysis in BUCKy assigns low concordance factors to most of the nodes separating major subgenera of *Heliconius*, which indicates that there exist multiple well-supported but conflicting gene tree topologies in this sample. Two of these nodes (*H. wallacei* and *N. aoede*; Fig. 2.3) are also only weakly supported by the *BEAST coalescent model, and correspond to the areas of high reticulation in the NeighborNet network, reflecting conflicting signals (Sup. Fig. S4). The same nodes are all assigned a posterior probability of one in the Bayesian supermatrix analysis (Fig. 2.1a), potentially leading to the erroneous conclusion that all the data point unequivocally to the inferred relations. The superiority of the MSC methods lies in the effective demonstration of incongruences, represented by lower support values or concordance factors assigned to the more difficult nodes (Belfiore et al. 2008).

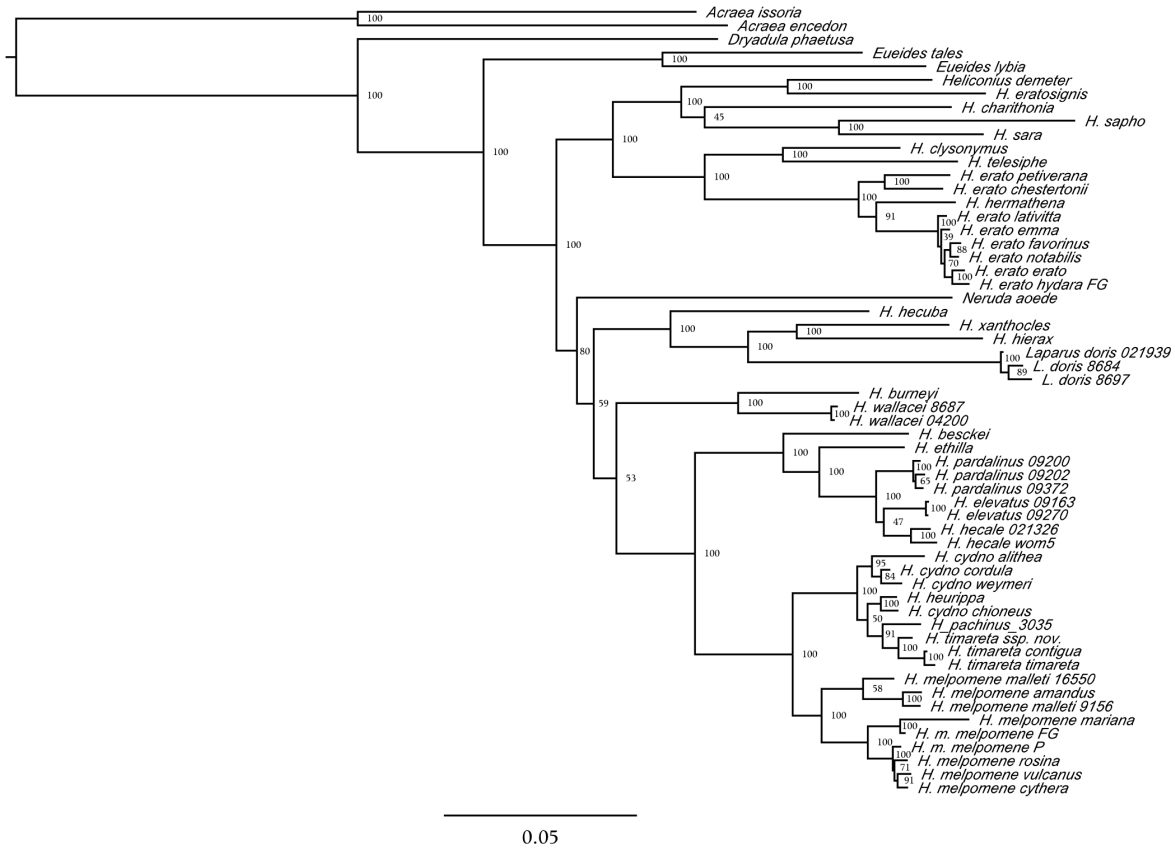


Figure 2.2. Whole mitochondrial maximum likelihood (RAxML) phylogeny of the genus *Heliconius.* Bootstrap support values indicated. Scale bar in units of substitution per site per million years.

Optimal sampling

My study highlights an important practical consideration in choosing the optimal analytical approach, where the requirements of the selected algorithm have to be reconciled with a realistic sampling of taxa. The final goal of testing hypotheses regarding the diversification dynamics of Heliconiini can be met only if the number of included taxa is maximized. Despite a substantial effort I only managed to secure single samples of the rare or geographically restricted species, and some of them are represented solely by historical

specimens with limited potential to generate extensive multilocus data. Six other species remain missing, as they are known only from a small number of specimens (*Neruda metis*, *Eueides emsleyi*) or found at low density in infrequently visited areas (*Philaethria andrei*, *P. browni*, *P. romeroi*, *H. lalitae*). Specimens of four of these were available to us, but poor preservation of material by desiccation made it impossible to extract reasonable quality DNA. Considering that the focal group is intensively studied and exceptionally well represented in research, museum and private collections due to its aesthetic appeal (Mallet et al. 2007), it would be considerably more challenging to obtain a complete sampling of many other groups. Another difficulty stems from the fact that the advanced coalescent techniques like BUCKy and *BEAST perform best with multiple samples per species, which should capture intraspecific diversity, and may require complete taxon coverage (Heled and Drummond 2010; Reid et al. 2014; Velasco and Steel 2014). Much of the uncertainty in the estimates can be attributed to missing data, which can negatively affect the estimation of both individual gene trees and the encompassing species tree (Wiens and Morrill 2011; Roure et al. 2013). When fitting a NeighborNet network, I found that although the percentage of data missing from the matrix does not explain all of the observed reticulation and the high delta score, it causes these parameters to increase, thus suggesting that the data completeness at each alignment position must be maximized to identify genuine incongruence. I observe that in many cases of biological interest it will be a formidable challenge to generate the ideal dataset that (i) has little missing data, (ii) comprises a large, genome-wide sample of loci, (iii) includes all taxa, and (iv) captures intraspecific variability. In case of Heliconiini, the supermatrix approach based on a limited number of markers (22) helped me to maximize taxonomic inclusiveness without compromising my ability to reconstruct a phylogeny in the light of conflicting biological signals.

Number of species and taxonomic implications

Heliconiini have been at the centre of the debate about species concepts and designation criteria, providing empirical evidence for the permeability of species barriers (Mallet et al. 2007; Kronforst et al. 2013; Martin et al. 2013; Nadeau et al. 2013). Species delimitation with the Poisson Tree Process based on relative branch lengths in the ML tree produced very surprising, although poorly supported results, strongly at odds with biological knowledge of Heliconiini. All the lineages in the *H. melpomene*/*H. cydno* group were lumped into a single Operational Taxonomic Unit (Sup. Fig. S7). This is despite extensive evidence for at least three species (*H. melpomene*, *H. cydno*, *H. timareta*), if not the currently recognized five (also including *H. pachinus* and *H. heurippa*): premating behavioural isolation (Merrill et al. 2011), host plant usage and habitat preference (Benson et al. 1975, Brown 1981, Merrill et al. 2013), morphology (Penz 1999) and genome-wide genetic structure (Nadeau et al. 2013, Arias et al. 2014). The small number of nucleotide differences between the biological species is consistent with interspecific gene flow, which led to a reduction in differentiation across the genome. Although automated species delineation is certainly a useful criterion for initial assessment of poorly understood taxa or for rapid inventory of entire communities (Zhang et al. 2013), I argue strongly that traditional taxonomies based on accumulation of extensive biological knowledge should take precedence in case of intensively researched taxa like Heliconiini. This is an important result, considering that the mitochondrial barcode region, although frequently used to delimit arthropod species at low cost, has also been shown to be insufficient to distinguish species in Ithomiini, a butterfly tribe co-mimicking Heliconiini (Elias et al. 2007).

Application of novel algorithms to Heliconiini data reveals a relatively stable topology and branch lengths, despite limited support (Fig. A2.1-3; Sup. Fig S2-7). I uphold the previous re-classification of the species in the genera *Neruda* (Hübner 1813) and *Laparus*

(Linnaeus 1771) as Heliconius based on their nested position (Beltrán et al. 2007). This placement is at odds with morphological evidence from adult and larval characters, which puts *Neruda* and *Laparus* together with *Eueides* (Penz 1999), suggesting convergent evolution of homoplasious morphological characters. Nevertheless, the molecular evidence is decisive and I thus synonymize *Laparus* syn. nov. and *Neruda* syn. nov. with *Heliconius*. I also conclude that the traditional naming of some *Heliconius* as "primitive" (e.g. Brown 1981) is phylogenetically unjustifiable and misleading. Instead, I propose to call these groups "the *wallacei* clade" (including *H. astraea*, *H. burneyi*, *H. egeria* and *H. wallacei*) and "the *doris* clade" (*H. doris*, *H. hecuba*, *H. hierax*, *H. xanthocles*).

The position of the enigmatic genus *Cethosia* remains unresolved, as it currently depends on the chosen method of analysis, and is unlikely to be established without a broad sampling of species using multiple markers. *Cethosia* has been variably considered to be either the only Old World representative of Heliconiini (Brown 1981; Penz and Peggie 2003; Beltrán et al. 2007), a genus of Acraeini (Penz 1999; Wahlberg et al. 2009), or possibly a distinct tribe (Müller and Beheregaray 2010). Establishing the systematic relations between Acraeini, *Cethosia* and Heliconiini is important for the study of Heliconiini macroevolution, as it will shed light on the dispersal route of Heliconiini into the Neotropics.

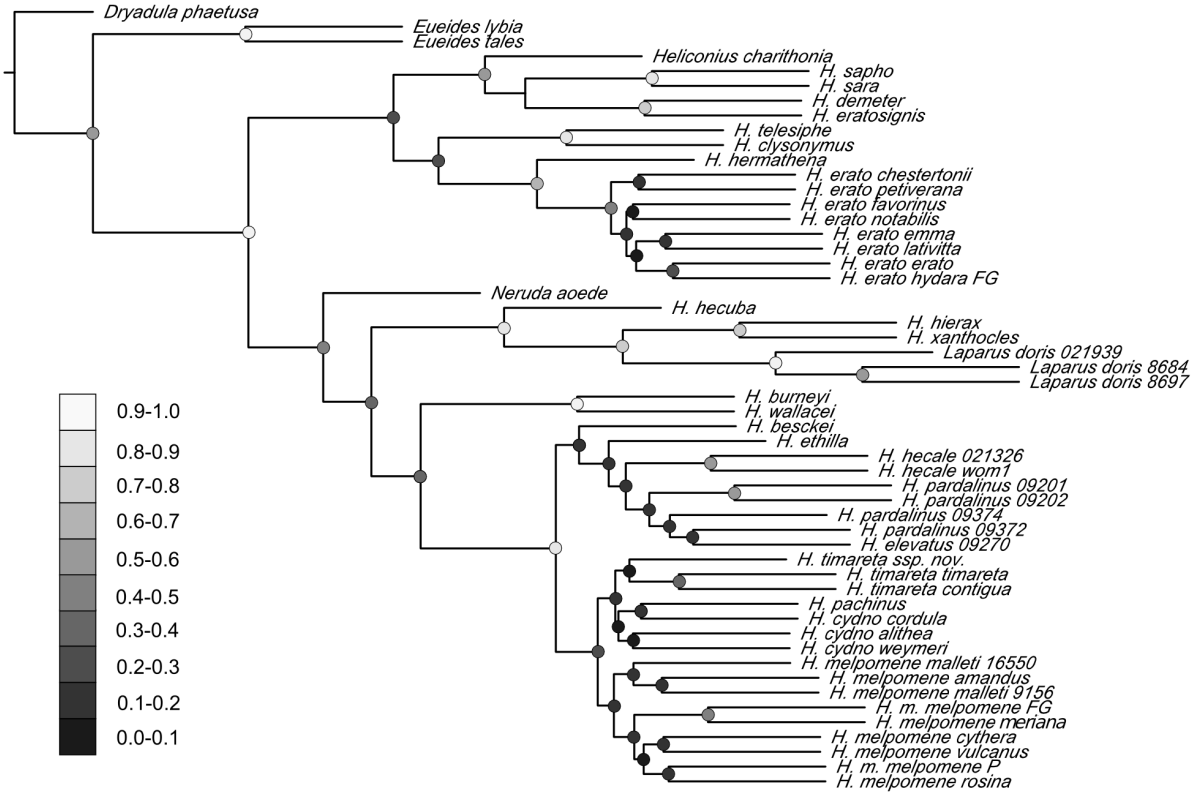


Figure 2.3. A phylogenetic hypothesis for *Heliconius* showing the extent of concordance between tree topologies of the 21 loci estimated by Bayesian concordance analysis in BUCKy. Dots indicate the Concordance Factor values for the nodes, with darker shades of grey corresponding to lower support values.

Divergence time estimates

My phylogeny of Heliconiini brings novel insights into the diversification dynamics of the clade. Although most age estimates for *Heliconius* agree with other studies, the deeper nodes are older than previously suggested (Table 1). There is little agreement on the dates above the species level, and the studies to date either suffer from insufficient taxon sampling (Pohl et al. 2009, Wahlberg et al. 2009, Cuthill and Charleston 2012; Table 3), or use markers unlikely to be informative above a relatively low level of divergence (Mallet et al. 2007). Fo

instance, the mean age of the split between *Heliconius* and *Agraulis* is estimated as 32 Ma in the study of Pohl et al. (2009), which includes only 3 species of Heliconiini; 26.5 Ma in Wahlberg et al. (2009), including one species per genus; or 23.8 Ma in the present study (Fig. 2.1a; Table 1). Conversely, Mallet et al. (2007) find the divergences between the genera of Heliconiini to be much younger than I propose, likely due to an effect of using a fast-evolving mitochondrial locus without partitioning (Brandley et al. 2011). Massardo and colleagues (2014) estimate broadly similar dates, but do not formally test alternative hypotheses of diversification and their conclusions are limited by incomplete sampling of the most diverse genera.

Our ability to correctly falsify the hypotheses regarding the diversification of Heliconiini hinges on having a nearly complete phylogeny, yet none of my MSC analyses consider as many species of Heliconiini as the Bayesian supermatrix estimate. I am confident that the values proposed by the supermatrix method can be trusted, as both the topology and the branch lengths are consistent with the results inferred by *BEAST based on a smaller dataset (Table 1; Sup. Fig. S3). I find that the ages of the deeper nodes and the length of terminal branches are not inflated by the supermatrix method in comparison to *BEAST, contrary to the predictions from simulations (Burbrink and Pyron 2011), and both methods infer similar mean age for the observed splits, for instance 11.8 Ma for the basal divergence of *Heliconius* into the *H. erato* and *H. melpomene* lineages, or around 3.5 Ma for *H. doris* and *H. xanthocles*. Hence I offer a new perspective on the dating of the Heliconiini radiation with a nearly complete set of divergence time estimates based on a well-resolved and supported tree.

Adaptive but not rapid radiation of Heliconius

Despite evolving many properties unique among the Lepidoptera (reviewed by Beltrán et al. 2007), *Heliconius* has not undergone a rapid adaptive radiation involving an explosive growth in the number of lineages, followed by a dramatic slowdown, as might be expected due to a sudden availability and filling of ecological niches (Pybus and Harvey 2000; Glor 2010; Moen and Morlon 2014). In fact, the rate of diversification has been increasing steadily for the last 11 Myr (Sup. Fig. S10). This result is consistent with a recent critique of the traditional prediction that large radiations should display a pattern of the initially high net diversification rate decreasing as the ecological niches fill up (Day et al. 2013). Such an expectation is reasonable for island radiations that constitute a large proportion of study cases to-date, but it does not necessarily apply to continental radiations in the tropics, where the scale and complexity of the ecosystems are likely to generate a number of suitable niches greatly exceeding even the cladogenetic potential of large radiations (Derryberry et al. 2011; Day et al. 2013). To date, steady diversification of a widely distributed taxon has been demonstrated in the 129 African *Synodontis* catfish (Day et al. 2013) and the 293 Neotropical Furnariidae ovenbirds (Derryberry et al. 2011), but the generality of this pattern remains unknown and requires further verification in other well-sampled large groups like Heliconiini.

Both Maximum Likelihood (DDD) and Bayesian (BAMM) approaches provide evidence for a substantial increases in speciation without changes in extinction rate, although I cannot completely exclude the possibility that extinction has not been modelled accurately and may have affected my findings. The first surge in speciation occurred early on the branch leading to the presently depauperate clade of *Podotricha telesiphe*, *P. judith* and the monotypic genera *Dryas* and *Dryadula*, and stands out against the low overall background rate of diversification (Fig. 2.1b, Sup. Fig. S10). More substantial increases are observed in the *Heliconius* and *Eueides* clade (*Heliconius* mean net diversification: 0.196 new lineages

per lineage per Myr; non-*Heliconius* rate: 0.101, Fig. 2.1b, Sup. Fig. S11). Bayesian modelling shows that the gains have been gradual and cannot be correlated with a single environmental factor, although the speciation of *Heliconius* and *Eueides* may have been stimulated by the second rapid stage of Andean orogeny ~12 Ma, which strongly changed the elevation gradients in the Central and Eastern sectors (Gregory-Wodzicki 2000; Blandin and Purser 2013). This was contemporaneous with climatic changes that affected the distribution of the rainforest (Lewis et al. 2007; Jaramillo et al. 2010), and followed shortly by the entrenchment of a major barrier, the Amazon, in its modern course 10 Ma (Hall and Harvey 2002; Hoorn et al. 2010). Nonetheless, my results do not strongly support such a correlation and highlight the dangers (Rabosky 2014) of the commonly used approach of selecting a single rate shift configuration (reviewed in Stadler 2013). For example, analysis of the Heliconiini data with DDD provided false confidence in a single distinct shift to a much higher rate of diversification. The lack of strong response by Heliconiini to the environmental perturbations is less surprising if we consider that intense changes on the continental scale have occurred nearly continuously from the start of Miocene (and thus Heliconiini) until the present. This has created a dynamic arena for constant evolution of new species and colonization of novel ecological niches (Derryberry et al. 2011), across a variety of environmental regimes, ecosystems and geological formations (Blandin and Purser 2013).

More puzzling is the gradual rise in the diversification rate of Heliconiini, clearly driven by the increased speciation rates of *Heliconius* (Fig. 2.1b, Sup. Fig. S10). One distinct possibility is that the extraordinary diversity of hundreds of mimetic patterns in the genus (see Fig. 2.1a for some examples) has facilitated speciation (Elias et al. 2008; Jiggins 2008b; Merrill et al. 2012) and stabilized extinction (Vamosi 2005), leading to a positive feedback of higher number of species on the diversification rate. However, testing any ideas on the macroevolutionary impact of the mimetic phenotypes will be difficult. For instance, Beltrán

attempted a reconstruction of the ancestral wing pattern by direct optimization on the phylogeny. This effort was fruitless, as the signal of ancestral morphological states is lost behind the existing intraspecific diversity of widely varied races and the confounding effects of pattern elements introgressing between species (M. Beltrán, PhD Thesis).

Some uncertainty surrounds the last few million years of evolution of Heliconiini. Early speculation regarding the drivers of speciation suggested diversification in allopatry, as the rainforest habitat occupied by most species in the group has undergone cycles of contraction and expansion in response to recent climatic variation (Turner 1965; Brown et al. 1974; Brown 1981; Sheppard et al. 1985; Brower 1994a). The hypothesis of vicariant cladogenesis has been subsequently criticized due to a lack of evidence for forest fragmentation in pollen core data (Colinvaux et al. 2000), lack of temporal concordance of divergence of distinct lineages (Dasmahapatra et al. 2010), and the likelihood of parapatric speciation (Mallet et al. 1998). The decrease in observed diversification over the last million years may be due to protracted speciation (Etienne and Haegeman 2012), i.e. my limited ability to delineate species in the assemblages of highly variable taxa such as *Heliconius.*

In conclusion, I present a taxonomically comprehensive phylogeny of a large continental radiation characterized by introgression and varying rate of diversification. Reticulate signals in phylogenetic data will increasingly have to be accounted for in study designs and could be common in other recent adaptive radiations, as recently reported for swordfish (Cui et al. 2013), mosquitoes (Crawford et al. 2014) or broomrape plants (Eaton and Ree 2013). Ignoring the possibility of reticulation no longer seems a viable assumption in phylogenetic analysis of recent adaptive radiations.

CHAPTER 3

A LARGE MOLECULAR PHYLOGENY OF PASSIFLORACEAE REVEALS EXTREMELY LABILE HOST PLANT RELATIONSHIPS AMONG HELICONIINI

At least a quarter of all known species on Earth are herbivorous insects (Janz et al. 2006), making insect-host plant interactions one of the key phenomena shaping ecology and evolution. Fifty years ago the seminal paper by Ehrlich and Raven (1964) brought the ubiquitous reciprocal interactions to the forefront of evolutionary theory and proposed the hypothesis of coevolution, whereby genetic changes in plants are caused by selective pressures from their insect predators, and in turn lead to evolutionary changes in herbivores. (Thompson 1999) characterised the Ehrlich and Raven model as “escape and radiate”, since the concept depends on the development of key innovations, which allow the lineage to escape selection and produce new diversity. Simulations show that coevolution is expected to increase diversity by accelerating the rate of speciation and promoting evolvability through enhancing selection on adaptations and counter-adaptations (Zaman et al. 2014). A study of 7500 insect families showed that host range is narrow at the family level and the interactions tend to be more specific in the tropics (Forister et al. 2015), but research at lower taxonomic levels exposes a wide spectrum of coevolved relations. For example, Nymphalini butterflies often revert from a specialised to a more generalist stage, which led to the alternative hypothesis of “oscillation” in the breadth of host range (Janz et al. 2001). A further prediction of this scenario is a rise in speciation rate (Janz et al. 2001; Janz et al. 2006).

Starting with the original work by Ehrlich and Raven, lepidoptera and their host plants have been a key system in which to test theoretical predictions (Thompson 1999; Janz et al. 2001, 2006; Ferrer-Paris et al. 2013). The Neotropical Heliconiini (Nymphalidae: Heliconiinae), revealingly dubbed passion vine butterflies, feed almost exclusively on the passion vines *Passiflora* and their sister genus *Dilkea* (Malpighiales: Passifloraceae) (Brown 1981; Janzen 1983). *Heliconius* and relatives have caught the attention and imagination of countless biologists, but the foundation for their fascinating biology lies in the intimate evolutionary relationship with *Passiflora.* Passion vine plants contain high concentrations of chemically varied cyanoglycosides with a rare cyclopentane ring moiety, which upon attack by caterpillars are converted by glycosidases to release cyanide (Spencer 1988). Very few insects have developed the ability to neutralise this defence, including some Coreidae (Hemiptera), flea beetles (Chrysomelidae), moths in the family and *Josia* (Dioptinae) and some Riodinid butterflies (Benson et al. 1975; Pemberton 1989). However, the most voracious and effective predators of *Passiflora* are Heliconiini, whose larvae not only overcome, but sequester the potent toxins, obtaining the compounds necessary for aposematism and thus Müllerian mimicry that drives the diversification of the tribe (Jiggins 2008). Although some toxins are produced endogenously by the butterflies (Chauhan et al. 2013), the majority are sequestered and the efficiency of uptake varies within the tribe, reaching the highest efficiency in *H. sara* and close relatives (Engler-Chaouat and Gilbert 2007), which additionally possess the ability to stop cyanide release altogether (Engler et al. 2000).

The 72 species of Heliconiini differ greatly in their host range, including both extreme generalists like *Agraulis vanillae* and complete specialists like *Heliconius telesiphe*, known to feed only on the eponymous *Passiflora telesiphe* (Benson et al. 1975; Knapp and Mallet 1998). Extensive field and laboratory studies of female oviposition and larval dietary choices

documented hundreds of associations (reviewed by Benson et al. 1975; Beccaloni et al. 2004; see Appendix: Bibliography) and a general mapping between major clades was proposed (Benson et al. 1975; Brown 1981). The smaller, older genera are obviously more generalist (Spencer 1988) and it is suggested that lineages within *Heliconius* show a tendency towards increasing specialisation, culminating in the oligophagy of the *H. sapho* clade (Benson et al. 1975). An impressive series of phytochemical studies in the 1980s (reviewed by Spencer 1988) catalogued rich variation of cyanoglycoside composition among *Passiflora* species, sufficient to recapitulate the basic taxonomy on the basis of chemical profiles alone, and discovered a weak correlation between diet preference and the "chemical class" of its preferred hosts.

Despite the abundance of observations, surprisingly little is known about the origin of the interactions between *Passiflora* and Heliconiini, and the overall importance of coevolution for speciation and adaptation in both groups. Few studies have used a phylogeny of Heliconiini to understand the patterns of plant usage (Brower 1997; Ossowski, MSc thesis) or toxin sequestration by the larvae (Engler-Chaouat and Gilbert 2007). The key obstacle for codivergence analyses has been the lack of a comprehensive *Passiflora* phylogeny (Brower 1997), leading most authors to rely on the dated monograph by Killip (1938), where a progression of *Passiflora* species is inferred from perceived primitive and derived aspects of morphology.

Passifloraceae is a diverse family of around 750 plants with a pan-tropical distribution, growing predominantly as vines, but also occasionally taking the form of small trees and shrubs (Krosnick et al. 2013). The genus *Passiflora* is by far the largest, with approximately 540 species, whereas the sister genus *Dilkea* contains only 12 (Ulmer and MacDougal 2004). During the Spanish *conquista* of South America the striking floral morphology commanded the attention of Jesuit friars, who found religious symbolism in the structure and named the

group after the passion of the Christ. Ever since, *Passiflora* have been praised for their extravagant aesthetics and popular as a decorative plant, resulting in an impressive variety of hybrid strains (Ulmer and MacDougal 2004). Some species are also cultivated for their fruit (see *P. edulis*), whereas others are increasingly recognised as potential indicators of ecosystem health in the Neotropics (Ocampo et al. 2010).

Feuillet and MacDougal update the taxonomy of *Passiflora* (2003) and several studies address the systematics of Passifloraceae based on morphological and molecular data (Muschner 2003; Yockteng and Nadot 2004). A series of papers by Krosnick and colleagues (Krosnick and Freudenstein 2005; Krosnick et al. 2009, 2013) highlights the history of the subgenera *Tetrapathea* and *Decaloba*. Muschner and colleagues use the relaxed molecular clock approach to estimate the dates of divergence between subgenera (2012). Tokuoka studies the relations between the genera of Passifloraceae (2012). These previous efforts are impressive, but unsatisfying. First, none of the studies attempts to address *Passiflora* systematics comprehensively, instead focusing on the higher levels (Tokuoka 2012), or a specific subgroup (e.g. *Decaloba*: Krosnick et al. 2013; *Cieca*: Porter-Utley 2014). Second, rather than consolidating and extending the available sequence data, some of the authors focused on phylogenetic marker discovery and evaluation (Yockteng and Nadot 2004; Krosnick et al. 2013). Third, the only chronogram of the family was calibrated with only one among many available fossils and unhelpfully published in a collapsed form (Muschner et al. 2012). Finally, botanists have treated the relation between *Passiflora* and the herbivores attacking it as an afterthought at best (Ulmer and MacDougal 2004; Krosnick et al. 2013). The first major goal of this work is therefore to synthesise the existing sequence and fossil data to produce a calibrated time-tree that includes the maximum possible number of Passifloraceae.

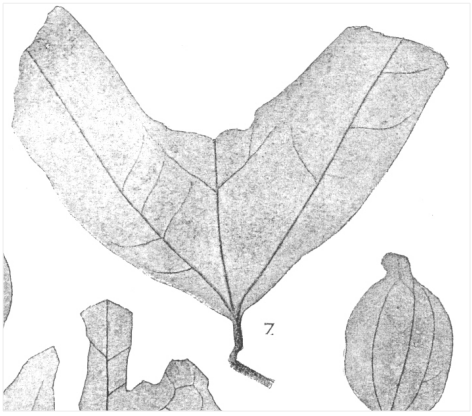


Figure 3.1. The leaf of *Passiflora*(?) *antiqua* (Newberry 1895) from the Raritan clays of New Jersey, dated at 99.6-93.5 MA. Although the trilobed shape, venation and curled stem closely resemble *Passiflora* synapomorphies, this fossil predates all others by at least 39 Myr.

The research on coevolution has hitherto focused on the macroevolutionary processes at the level of families (Janz et al. 2001; McKenna et al. 2009; Ferrer-Paris et al. 2013; Forister et al. 2015), but broad studies generalising across a range of well-understood specific interactions to dissect the course of coevolution and alternative processes in a large and diverse radiation remain a rarity (but see Cruaud et al. 2012)). The Heliconiini-Passifloraceae system is a perfect study case in which to combine the understanding of specific processes with a nearly complete picture of the pattern. In this chapter I catalogue the results from decades of research into the ecological interactions, fill in the major gaps in our understanding of the evolutionary dynamics of Passifloraceae and use novel phylogenetic hypotheses to answer the following questions:

1. Is there evidence for an escape and radiate scenario with increasing specialisation of the more recent lineages within Heliconiini (Benson et al. 1975; Fordyce 2010)? Are some lineages more generalist and capable of host switching?
2. Is the diversity within clades of Heliconiini determined by the diversity of their respective host plants (Janz et al. 2006)?

3. What is the evolutionary pattern of herbivore pressure on the passion vines and how does it shape the evolution of their defences (Benson et al. 1975)? Are some species more likely to be attacked?
4. Is there a signal of cospeciation between passion butterflies and their hosts (Ehrlich and Raven 1964)? Do subgenera of Heliconiini choose specific subclades of *Passiflora* (Benson et al. 1975)?
5. Are the two radiations contemporaneous, as inferred for other groups such as beetles (McKenna et al. 2009)?

METHODS

Inference of the Passifloraceae phylogeny

Taxonomic notes: 1. The italicised name *Passiflora* has been used for the entire genus; one of the two large subgenera alongside *Decaloba*; or a supersection within the subgenus. I refer to the genus as "passion vines" and reserve the name *Passiflora* for the subgenus unless specified otherwise. 2. The name Passifloraceae is used *sensu lato*, including the formerly separate family Turneraceae (Tokuoka 2012).

The phylogenetic relations between Passifloraceae were inferred entirely from DNA data published in GenBank and TreeBase as of 2013 (Kaellersjoe et al. 1997; Savolainen et al. 2000; Muschner 2003; Yockteng and Nadot 2004; Alford 2005; Krosnick and Freudenstein 2005; Krosnick 2006; Wurdack and Davis 2009; Muschner et al. 2012; Tokuoka 2012; Krosnick et al. 2013). I chose seven plastid (*atpB, matK, ncpGS, ndhF, rbcL, trnLT, trnLF*), three mitochondrial (*nad1, nad5, rps4*) and two nuclear (*ITS, cytGS*) genes, based on the number of sequences available, whether unique species were included, and if alignment across species was possible (Table 3.1). Sequences were aligned in CodonCode v. 4

(CodonCode Corporation 2012) and MUSCLE (Edgar 2004), inspected for misalignment in Mesquite v. 2.75 (Maddison and Maddison 2011) and trimmed at the ends. The genes *ITS* and *trnLF* were divided into two and three partitions respectively, as various authors sequenced different sections of these loci. The best-fitting models of substitution were inferred under the BIC in jModelTest v. 2 (Posada 2008), and individual gene trees were estimated in PhyML v. 3 (Guindon et al. 2010). Properties of the partitions are listed in Table 3.1. Maximum Likelihood trees for all supermatrices were estimated in RAxML v. 8 (Stamatakis 2006) with 200 bootstrap replicates, under a separate GTR+Γ model for each partition.

DNA data were available for 273 species in the genera *Passiflora* and *Dilkea*, including 99 heliconian host taxa. Phylogenetic studies based only on published data typically suffer from the tradeoff between taxonomic inclusiveness and quality of the alignment. Inclusion of more taxa improves modelling and breaks up long branches (Heath et al. 2008), but many species are only represented by one or a few markers and the resulting matrix of all available sequences is very heterogeneous and gappy, which could lead to inference errors (Wiens 2003; Lemmon et al. 2009; Wiens and Morrill 2011). Properties of molecular clock rate estimates from sparse matrices are especially poorly understood (Filipski et al. 2014). I controlled for these problems in three steps.

1. I considered all 273 species of *Passiflora* and *Dilkea* for which data was available, as well as 37 Old World Passifloraceae and seven *Malesherbia* outgroups giving a **total of 310** species (Wurdack and Davis 2009; Soltis et al. 2011). However with this dataset the ML tree inference suffered from multiple long branch artefacts (e.g. *Basananthe* nested in *Passiflora*) in cases where distantly related species were represented by the same rare markers (Fig. S3.1).

2. I eliminated the “rogue taxa” (Thomson and Shaffer 2010): 14 outgroups with exceptionally incomplete data and enforced the well-established monophyly of the ingroup

genera *Passiflora*, *Hollrungia*, *Tetrastylis* and *Tetrapathea* (Krosnick et al. 2009, 2013; Tokuoka 2012) (**297 species total**). To investigate the impact of partitioning a sparse matrix on model inference, an optimal partitioning scheme was selected with PartitionFinder (Lanfear et al. 2012). ML trees were estimated from matrices: (a) without partitions; (b) partitioned by organelle (3 sections); (c) partitioned by gene (15); (d) partitioned optimally (15). Likelihood of the trees was compared with an hLRT. Support was also expressed as the RAxML Total Internode Certainty (TCA) measure, which reflects the frequency of observed bipartitions in relation to all alternatives found in the bootstrap trees (Salichos et al. 2014). Partition schemes (a), (c) and (d) were also implemented in BEAST (details below) and their likelihoods were compared using AICM (Baele et al. 2012) in Tracer (Rambaut et al. 2014) with 1000 bootstraps.

3. To minimise the modelling problems caused by the sparse matrix I considered only 99 species attacked by *Heliconius*, 15 other passion vines with exceptionally good data to increase the overlap between partitions (Thomson and Schaffer 2009; Filipski et al. 2014), and all outgroups (**157 species**).

Estimation of divergence times

Unlike the Heliconiini butterflies, both Passifloraceae and Malesherbiaceae have a good fossil record that can be used to calibrate the molecular clock, but none of the previous studies present the palaeontological evidence exhaustively (Krosnick 2006; Muschner et al. 2012). An extensive literature search produced 12 records (Table 3.2; Appendix Bibliography). However, not all fossils are created equal: the issue of how to choose and incorporate reliable calibrations is a subject of intense debate (Pulquério and Nichols 2007; Forest 2009; Sauquet et al. 2012; Heath et al. 2014). A fossil is useful only if it can be assigned to a clade with reasonable certainty, and the stratum it comes from can be dated

(Parham et al. 2012). Many of the supposed Passifloraceae are partial fossils with few reliable synapomorphies (Table 3.1). For instance, the frequently reported pollen of the family has few external characters for identification (Silvério and de Araujo Mariath 2013; Tim Upson, *pers. comm.*). The oldest reference (Newberry 1895) presents a fossil resembling the modern *P. biflora* (Fig. 3.1), but the similarity may simply be a taphonomic artefact. Wherever possible, I examined the original published drawings of the material and checked the age of the stratigraphic units against independent sources (Table 3.1).

I chose three reliable calibration points (Table 3.1). The age of the most common recent ancestor (MRCA, crown calibration) of the genus *Passiflora* was set to at least 16.0 MA (exponential distribution: mean=0.5, offset=16.0), based on multiple fruit, seed and pollen fossils from Central Europe (Table 3.1) (Mai 1967; Gregor 1978). The minimum age of Passifloraceae was constrained to 33.9 MA based on *Passifloraephyllum krauseli* leaves from the Upper Eocene of Hungary (Rasky 1960). The crown age of Passifloraceae *sensu lato* (Tokuoka 2012) was modelled as normally distributed with a mean of 48.5 MA of and a standard deviation of 6.5 MA, as inferred in a phylogenomic study of Malpighiales (Davis et al. 2005; Xi et al. 2012). The maximum age of divergence from the sister family Malesherbiaceae was bounded at 110.0 MA based on habitat availability (Xi et al. 2012).

Taxon	Location	Rock	Epoch	Age (MA)	Material	Synapomorphy	Reliability	Fossil ref.	Stratum ref.
Passiflora kirchheimeri	Oberpfaelz, Germany	clay, coal	Burdigalian, Lower Miocene	20.4-16.0	fruit and seed		high	Gregor 1978	fossilworks.org
Passiflora OR Turnera	East Germany		Badenian, Mid-Miocene	16.5-13.0	seed	pitting of the seed wall	high	Mai 1967	fossilworks.org
Passifloraephyllum krauselii	Budapest-Obuda, Hungary	coral-algal marl	Priabonian, Upper Eocene	37.2-33.9	leaves	"general morphology"	high	Rasky 1960	Kazmer 1985
Passiflora sp.	Paraje Solo, Veracruz-Lave, Mexico	lignite	Upper Miocene	11.6-5.3	pollen		low	Graham 1975; Porter-Utley 2003	Vokes 1970
Passiflora kirchheimeri	Cukurovo, Bulgaria	coal-bearing	Mid-Miocene	16.0-11.6	pollen (?)		low	Palamarevet al. 1971, 2005	
Passiflora heizmannii	Sandelzhausen, Bavaria, Germany		Mid-Miocene	16.0-11.6	?		low	Gregor 1982	
Passiflora sp.	Turow, Lower Silesia, Poland		Burdigalian, Lower Miocene	23.0-16.0	?		low	Czeczott and Skirgiello 1965	Holy et al. 2012
Passiflora antiqua	Raritan Formation, Woodbridge, New Jersey	clayey slit	Cenomanian, Cretaceous	99.6-93.5	leaf	venation, leaf shape	low	Newberry 1895	Christopher 1979
Stephanocolpites sp. (Passifloraceae)	Colombia		Upper Paleocene	55.8-58.7	pollen		low	van der Hammen 1954; Jaramillo et al. 2010	
Turneraceae?	Ogwashi-Asabe Formation, SE Nigeria	lignite, sandstone	Middle Eocene	47.8-38.0	pollen		low	Jan du Chene et al. 1978	Bassey and Eminue 2012
Passifloraceae? (may be Proteaceae)	Taratu Formation, Livingston, N. Zealand		Early-Mid Eocene	54.0-38.0	leaf fragment	leaf serration	low	Pole 1994	Carter 1988
related to Passifloraceae	Nigeria		Maastrichtian, Cretaceous	70.6-65.5	seeds		low	Chesters 1955	

Table 3.1. Only three fossils of Passifloraceae can be confidently classified and independently dated. Multiple Eastern European floras date the minimum age of *Passiflora* to at least 16.0 MA. Known fossils were identified through a systematic literature and database search. Fossil and stratum references listed in Appendix: Bibliography.

Times of divergence between the **157** or the **297** species were inferred in BEAST v. 1.8 (Drummond et al. 2012). A separate substitution model was set for each gene partition. To avoid overparametrisation an uncorrelated relaxed molecular clock model was implemented for each organellar partition (plastid, mitochondrial, nuclear) under the gamma prior with scale set to 0.001. I used an ML starting tree, incomplete sampling birth-death tree prior (Stadler 2009), and a constraint on the monophyly of the ingroup. I executed three independent runs of $100*10^6$ rounds of MCMC each, sampling the chain every 10^4 cycles. I evaluated convergence in Tracer v. 1.6 (Rambaut et al. 2014), discarded the first $2*10^7$ cycles as burnin and produced the Maximum Clade Credibility tree in TreeAnnotator (Drummond et al. 2012). xml files of the data and models for the 297 taxa and is included in the Appendix.

Geographical and temporal patterns

The dispersal patterns of Passifloraceae were investigated by Ancestral State Reconstruction (ASR) of geographical ranges on the **157** species tree under the MP criterion in Mesquite. Present-day distributions at the sub-continental scale were obtained from literature (Ulmer and MacDougal 2004; Krosnick 2006; Muschner et al. 2012; Tokuoka 2012) and biodiversity information repositories (Global Biodiversity Information Facility 2015; Encyclopedia of Life 2015).

Temporal dynamics of the family were inferred based on the **297** taxa MCC chronogram. An LTT plot was drawn in the R package *ape* (Paradis 2004).Changes in the rate of diversification in *Passiflora* were modelled by Bayesian Analysis of Mixture Models (BAMM) and visualised with R *BAMMtools* (Rabosky et al. 2014a, 2014b). I assumed a sampling fraction of 0.5 (269/540 species: Krosnick et al. 2013) and executed two independent runs of 10^6 steps. Alternative models were compared using Bayes Factors.

Tests of cospeciation

Data on the *Heliconius* larval feeding preferences were collected in an extensive literature search, with most of the records coming from the classic summary by Benson and colleagues (Benson et al. 1975), and the general catalogue by Beccaloni et al. (2008). I considered only observations from the wild, as the dietary specialisation is often behavioural and some captive Heliconiini can be induced to feed on passion vines that they do not consume in the wild (Silva et al. 2014). The 12 species of *Dilkea* are usually not distinguished in *Heliconius* literature and were therefore treated as a single taxon. The 502 interactions are fully referenced in Appendix Data.

Robust estimates of butterfly and host plant phylogenies make it possible to test and visualise cospeciation with host-parasite methods. Two distinct approaches were compared. Parafit is an established algorithm that solves "the fourth corner problem": the general host-parasite association between the two trees is inferred from (i) matrices of binary interaction data (e.g. *Passiflora* is/is not predated by *Heliconius*), (ii) pairwise distances between hosts and (iii) pairwise distances between the butterflies (Legendre et al. 2002). The results describe whether the two topologies are non-random with respect to one another, and which individual associations are likely to be due to coevolution. The algorithm was run through the CopyCat v. 1.14 interface using the AxParafit optimisation (Meier-Kolthoff et al. 2007; Stamatakis et al. 2007). Significance of the associations was tested with 10^5 permutations.

Jane is a complementary event-based method, which uses a genetic algorithm to find the optimal mapping of the parasite tree onto the host tree, given the associations between tips and an explicit cost matrix for various events, such as duplications or host switches (Conow et al. 2010; Cruaud et al. 2012). The result is a model of parasite evolution including specific changes. Jane v. 4 was run with the default cost matrix, as well as with equal cost for all event types. I ran 30 generations of the genetic algorithm with a population size of 600

(R. Liebeskind-Hadas, unpublished). The significance was tested against a null distribution of 10^4 randomised replicates.

Coevolution

I hypothesised that the more recently derived clades of Heliconiini tend to be more oligotrophic (Benson et al. 1975; Brown 1981). To visualise the trends, I optimised the number of host plants on the phylogeny of Heliconiini under the MP criterion in Mesquite v.2.75. In parallel, I mapped the number of Heliconiini predators onto the Passifloraceae phylogeny. The differences in the mean number of butterfly herbivores between the two major subgenera (*Passiflora* and *Decaloba*), as well as between 22 smaller sections of *Passiflora* (Ulmer and MacDougal 2004; Krosnick et al. 2013), were tested with the Kruskal-Wallis rank sum test. To visualise the dietary similarities between butterflies, I coded the preferences as binary characters and generated an MDS plot with the *cmdscale* function in R.

The ParaFit analysis determined how many of the dietary preferences exhibited by a butterfly species are due to coevolution. For each species I calculated a ratio of such coeveolved associations to the total number of host plants used, which I call the "Conservatism Ratio" (CR), as a high value indicates that the diet of the species is strongly predicted by phylogeny.

Conservatism Ratio = (# coevolved plant associations) / (# host plants)

Two ratios were computed, based on the associations considered coevolved at the p=0.01 or the more liberal p=0.05 significance level. A phylogenetic ANOVA in the R package *GEIGER* (Harmon et al. 2008) was used to test the differences in host number and *CR* between the major clades of Heliconiini. Blomberg's K was computed with 10^4 permutations to assess the phylogenetic signal in the number of hosts per butterfly species under a Brownian motion null-model in *picante* (Blomberg et al. 2003; Kembel et al. 2010).

RESULTS

Genetic, fossil and ecological data

Public sequence repositories contained DNA data for 270 out of 540 species of passion vines, including 254 from the Neotropics; three out of 12 *Dilkea*; and 37 other Passifloraceae from 23 genera. 11 loci from all three genomes were chosen, including unusual genes like *cytGS*. The long *trnLF* and *ITS* alignments had to be further subdivided to account for amplification with different primer pairs. Most partitions were relatively long (399-1736 bp), but the number of species across partitions varies from 37 to 181, and individual genes differ in variability, from classic deep phylogeny markers (e.g. *matK),* to ones resolving the recent divergences (e.g. *cytGS*; Table 3.2).

12 potential Passifloraceae fossils are known (Table 3.1), three of which can be reliably identified and dated. Multiple plant parts of varied quality were preserved in several Miocene strata of Central Europe (Mai 1967; Rasky 1975; Appendix: Bibliography), which together show that the genus *Passiflora* was certainly present at least 16.0 MA. Pollen finds are reported from the Americas, Africa and Europe. Two fossils date to Eocene and one is known from Cretaceous deposits, but the morphological characters may not be enough for a correct identification (Fig. 3.1; Table 3.1).

651 Passifloraceae-Heliconiini associations are known from the wild (Appendix Data: Associations Table; Appendix Bibliography). Of those, 502 are included in the codivergence analysis, i.e. both the plant and the butterfly are found in the phylogenies. No sequence data is available for 61 out of 160 passion vines attacked by Heliconiini, which I suspect to be uncommon and infrequently used species.

Gene	Type	Genome	Taxa	bp	Model	Original source
rbcL	exon	plastid	181	1304	HKY+Γ	Kjaellersoe et al. 1997, Savolainen et al. 2000, Wurdack and Davis 2009, Tokuoka 2011, Muschner et al. 2012
rps4	exon	mitochondrial	132	617	GTR+Γ	Muschner 2003, Muschner et al. 2012
trnLT	intron	plastid	41	605	TPM1uf	Muschner et al. 2012
trnLF (part1)	intergenic spacer	plastid	123	1093	TPM1uf+Γ	Muschner 2003, Alford 2005, Muschner et al. 2012, Krosnick et al. 2013
trnLF (part2)	intergenic spacer	plastid	152	399	HKY+Γ	Muschner 2003, Alford 2005, Muschner et al. 2012, Krosnick et al. 2013
trnLF (part3)	intergenic spacer	plastid	53	660	TPM1uf+Γ	Muschner 2003, Alford 2005, Muschner et al. 2012, Krosnick et al. 2013
nad1	intron	mitochondrial	89	1686	TPM1+Γ	Wurdack and Davis 2009, Muschner et al. 2012
nad5	intron	mitochondrial	88	1530	TPM1+Γ	Muschner et al. 2012
matK	exon	plastid	68	1239	GTR+Γ	Wurdack and Davis 2009, Tokuoka 2011
atpB	exon	plastid	48	1469	GTR+Γ	Wurdack and Davis 2009, Tokuoka 2012
ITS part1	ribosomal gene	nuclear	46	1736	TrNef+I	Wurdack and Davis 2009, Tokuoka 2011, Muschner et al. 2012, Krosnick et al. 2013
ITS part2	ribosomal gene	nuclear	*157*	727	TRNef+Γ	Wurdack and Davis 2009, Tokuoka 2011, Muschner et al. 2012, Krosnick et al. 2013
ncpGS	exon+intron	plastid	130	760	HKY+Γ	Krosnick et al. 2013
cytGS	exon+intron	nuclear	37	805	HKY	Krosnick 2006
ndhF	exon+intron	plastid	132	751	GTR+Γ	Krosnick et al. 2013
	SUPERMATRIX		310	13581	GTR+Γ	

Table 3.2. DNA loci used to estimate the Passifloraceae phylogeny. Taxa numbers and models reported for the largest dataset (310 species).

Passifloraceae phylogeny

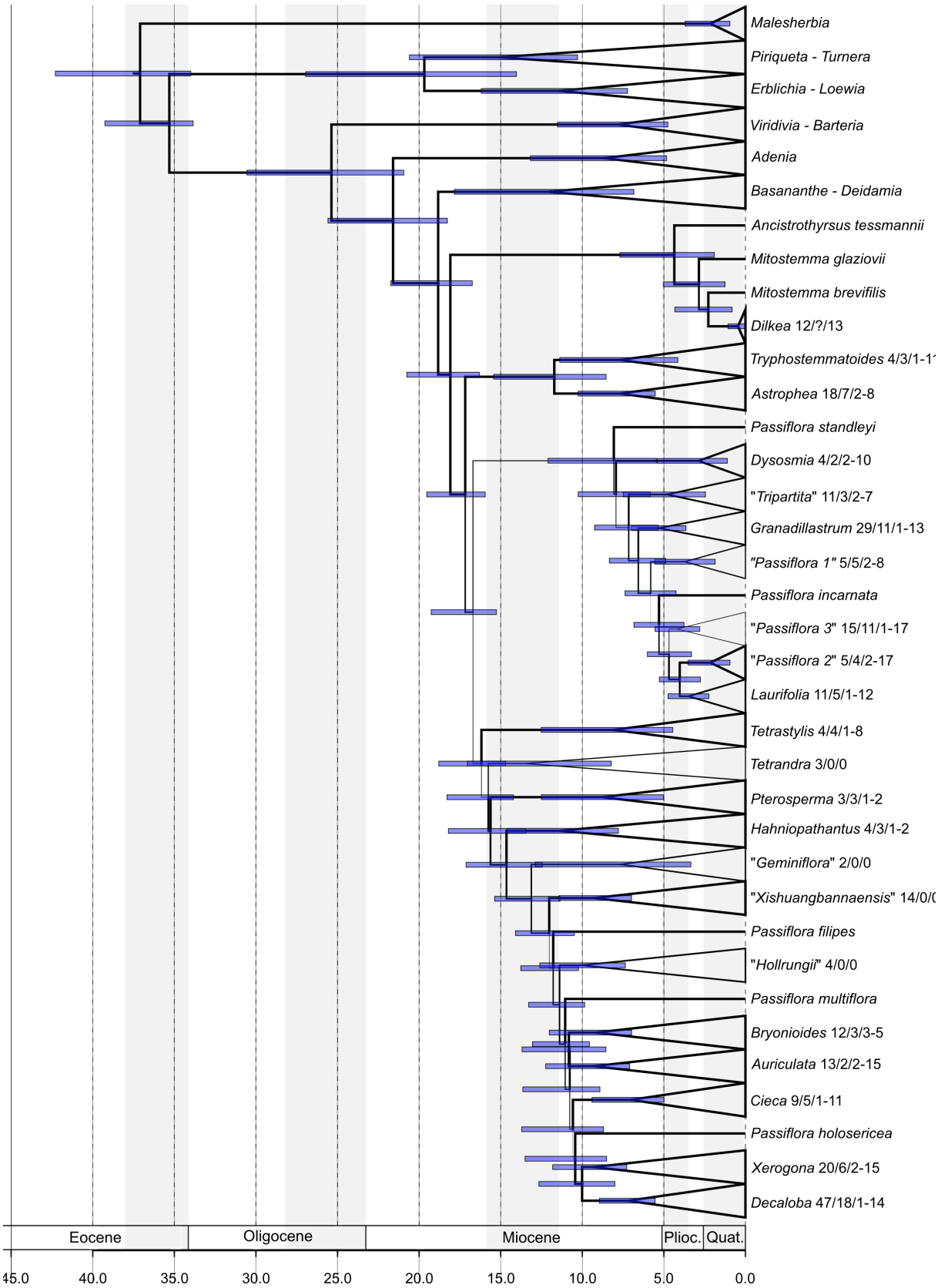
The inferred topologies of Passifloraceae are robust (Fig. 3.2) and nearly identical between the optimality criteria. Bayesian chronograms were used downstream, as I considered them more informative. My results are entirely consistent with previous studies at the level of genera (Tokuoka 2012) and subgenera (Muschner et al. 2012), as well as broadly similar to the species relations inferred by Krosnick and colleagues (2013). Notable differences include the placement of the three species related to *Tetrastylis ovalis* closer to the subgenus *Decaloba* than *Passiflora*. The order of branching between several small supersections or sections (*Multiflora*, *Cieca*, *Bryonioides*) and some old species (*P. multiflora*, *P. hollrungii*) is also different (Fig. 3.2), albeit the conflicted branches were not well-supported in the previous study either (Krosnick et al. 2013). My results and other work reveal the incompatibility of molecular systematics with morphological taxonomy (MacDougal and Ulmer 2004), as many large supersections are polyphyletic, e.g. *Passiflora* and *Auriculata* (Krosnick et al. 2013). As a taxonomic revision is beyond the scope of this work, I use *ad hoc* descriptions that correspond to empirical clades (e.g. supersections *Passiflora 1*, *2* and *3*).

A sampling error caused the Bayesian divergence time estimates to be much higher when based on the 157 rather than the 297 taxa supermatrix. For instance, ages of the root and the *Dilkea/Passiflora* split are estimated at respectively 51.6 (95% HPD: 38.6-65.1) MA and 32.2 (26.6-41.8) MA with 157 taxa, but only 37.2 (32.0-42.3) MA and 18.1 (16.4-20.8) MA for 297 taxa (Appendix: BEAST xml and BEAST chronograms). Even older dates are estimated in a Bayesian analysis of just 106 species (Muschner et al. 2012). Variability of an order of magnitude is documented among various dating schemes for the beech genus *Nothofagus*, but those differences are driven by priors (Sauquet et al. 2012), and here all priors were kept constant. What else could cause the differences? (1) The PartitionFinder scheme for partitioning the supermatrix was selected by the hierarchical Likelihood Ratio Test

(p<0.0001), but the Bayesian test favoured a simpler partitioning by gene (ΔAICM=3571.5). Regardless, the phylogeny estimates are robust to changes in partitioning schemes, which yield nearly identical topologies, support values (TCA) and divergence estimates (Table 3.3). (2) The disparity in split times could also be due to differences in the matrix composition, leading to errors in topology (Heath et al. 2008; Wiens and Morrill 2011) or in the clock models (Filipski et al. 2014). However, both matrices have a similar structure: 70.3% missing data and 33.7% pairwise sequence identity (157 taxa); *versus* 74.3% missing, 38.4% identity (297 taxa). Furthermore, all sections (*sensu* MacDougal and Ulmer 2004) overlap at multiple loci, in which case much greater amounts of missing data could be tolerated (Filipski 2014), and the gaps are favourably distributed in blocks (Fig. S3.1) (Wiens and Morrill 2011). (3) Thus the difference in the number of species sampled remains as the primary reason for the disparate time estimates.

Other differences between the two trees are minor. The 297 species tree has lower posterior probabilities along the stem (Fig. 3.2; Appendix), but most of the same branches are recovered with higher support in the smaller tree. The two phylogenies differ at only 18 branches, 16 of which are unsupported. The age estimates are also statistically compatible as the 95% HPDs around split times overlap, although the larger tree has much lower error in time estimates (full trees in Appendix). The phylogeny based on more extensive sampling should be preferred as more likely to be correct (Heath et al. 2008).

Figure 3.2. Bayesian chronogram of 297 Passifloraceae estimated in BEAST with tips collapsed into major clades (fully resolved tree can be found in the Appendix). Numbers at the tips indicate: number of tips in the clade/number attacked by Heliconiini/range of the number of butterflies hosted. Bars represent the 95% HPD around the mean age of clade. Axis in MA. Branch thickness is proportional to posterior probability.


Malesherbia
Piriqueta - Turnera
Erblichia - Loewia
Viridivia - Barteria
Adenia
Basananthe - Deidamia
Ancistrothyrsus tessmannii
Mitostemma glaziovii
Mitostemma brevifilis
Dilkea 12/?/13
Tryphostemmatoides 4/3/1-1
Astrophea 18/7/2-8
Passiflora standleyi
Dysosmia 4/2/2-10
"Tripartita" 11/3/2-7
Granadillastrum 29/11/1-13
"Passiflora 1" 5/5/2-8
Passiflora incarnata
"Passiflora 3" 15/11/1-17
"Passiflora 2" 5/4/2-17
Laurifolia 11/5/1-12
Tetrastylis 4/4/1-8
Tetrandra 3/0/0
Pterosperma 3/3/1-2
Hahniopathantus 4/3/1-2
"Geminiflora" 2/0/0
"Xishuangbannaensis" 14/0/(
Passiflora filipes
"Hollrungii" 4/0/0
Passiflora multiflora
Bryonioides 12/3/3-5
Auriculata 13/2/2-15
Cieca 9/5/1-11
Passiflora holosericea
Xerogona 20/6/2-15
Decaloba 47/18/1-14
Eocene
Oligocene
Miocene
Plioc.
Quat.
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0

Partitioning	-lnL	Parameters (d.f.)	hLRT *p*-value	ΔAICM	RelTCA	*Passiflora* age (MA)
None	109629.48	10 (0)	n/a	0	0.4168	37.4, 17.5
Organelles	109474.93	28 (18)	<0.0001	n/a	0.4212	n/a
By gene	107830.08	136 (108)	<0.0001	3571.5	0.4003	37.6, 17.4
Optimal	107109.83	136 (0)	<0.0001	2724.9	0.4125	38.0, 17.4

Table 3.3. Alignment partitioning does not affect the inference of Passifloraceae phylogeny. A scheme selected by PartitionFinder scores the highest, but the support and split times are almost unchanged. Alignment partitioning schemes for the **297** species were evaluated based on a hLRT and the total relative Tree Certainty (TCA) of the ML trees, and the AICM calculated from a Bayesian posterior distribution. Node ages estimated in BEAST.

Evolution of Passiflora

Dating with two calibrations from multiple fossil floras (Table 3.1) and two broad secondary constraints (Xi et al. 2012) shows that Passifloraceae *sensu lato* diverged from Malesherbiaceae 37.2 MA (95% HPD: 34.0-42.23) and the divergence within the family followed shortly after at 35.4 MA (33.9-39.3). The MRCA of *Heliconius* hosts *Passiflora* and *Dilkea* appeared 18.9 MA (16.8-21.8) (Fig. 3.2). Supersections and sections of *Passiflora* diversified rapidly: 13 lineages appeared in the 15-10 MA window, which is likely why the corresponding branches are poorly supported here and in other work (Krosnick et al. 2013). My estimates are much lower than those inferred from a different set of calibrations, such as the root at ~80 MA (Muschner et al. 2012).

My results uniquely combine wide taxonomic and geographic scope to facilitate phylogeographic inference about the family. The outgroup *Malesherbia* is found in the Americas, but the grade of taxa leading to *Passiflora* contains a number of Old World lineages (Fig. S3.2). Ancestral state reconstruction shows that the *Stapfiella, Tricliceras, Loewia* and

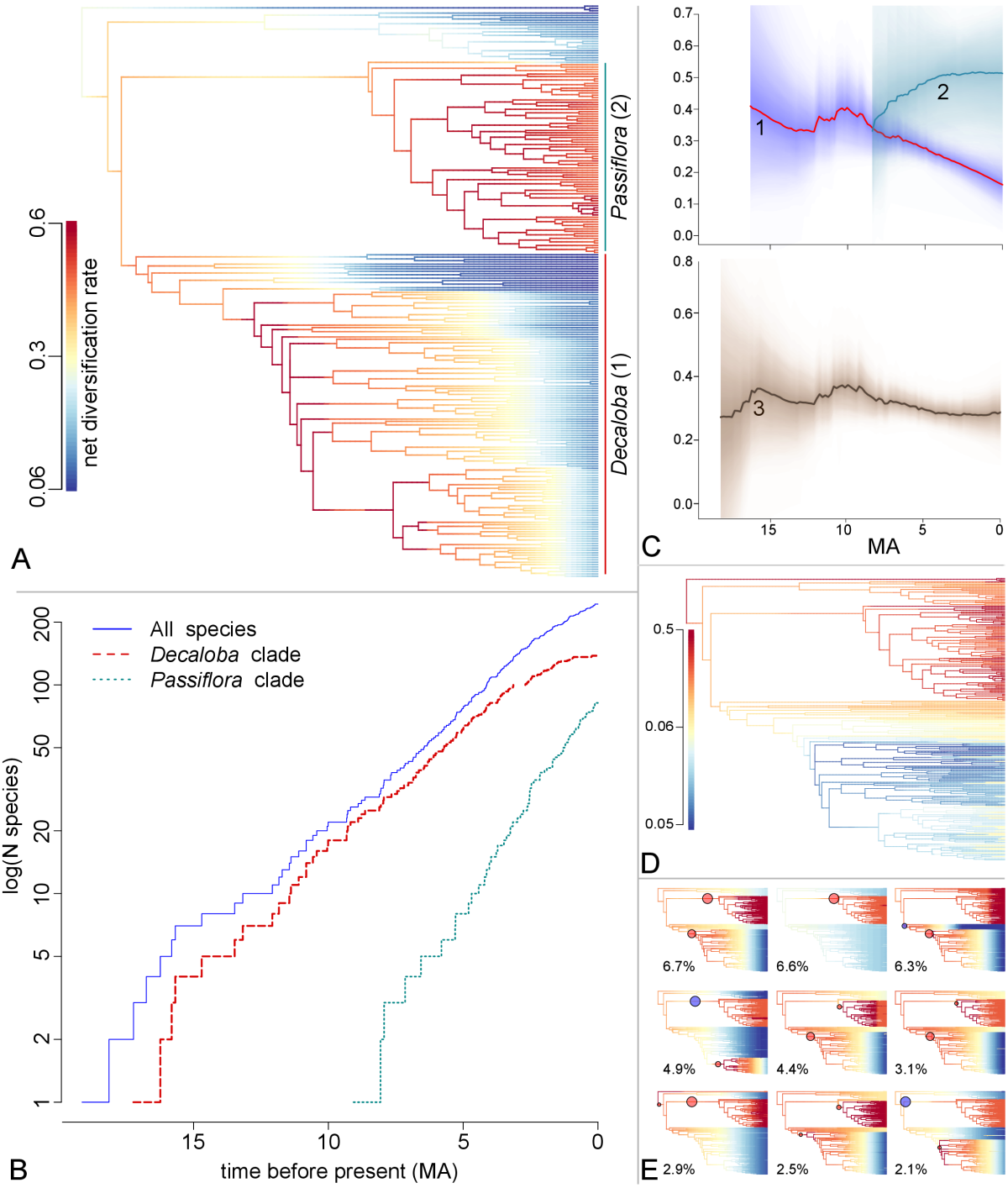
Streptopetalum lineage shared a common ancestor with *Mathurina* in South America, but around 13 MA colonised Africa. Their sister lineage invaded Africa earlier (28-32 MA), but later produced the Neotropical group of *Ancistrothyrsus*, *Mitostemma*, *Dilkea* and *Passiflora* (21-22 MA). Australasia was colonised independently twice, by *Tetrapathea* (*Passiflora*) *tetrandra* and *Passiflora xishuangbannaensis* clades. The Caribbean lineages have several independent origins (Fig. S3.2).

To understand the dynamics of passion vine diversification I modelled speciation and extinction rates on the MCC chronogram of 50% of the extant species using BAMM (Rabosky et al. 2014a). I estimated that the number of species in the genus increased exponentially, but the rate slowed down in the subgenus *Decaloba* after the upper Miocene, and was compensated for by the rapid growth in the subgenus *Passiflora* (Fig. 3.3.A&B). The net diversification rate peaked twice, at 16 and 11 MA, as sections diverged within the subgenus *Decaloba* (Fig. 3.3.C3). It has to be noted that the estimates are uncertain and the error around the mean increases for more ancient events, reaching as much as 40% of the rate (Fig. 3.3C). A closer look reveals different processes in the major subgenera, since around 7.5 MA the rate of diversification in *Decaloba* drops below the average level of 0.35, but new lineages of *Passiflora* start to appear rapidly (Fig. 3.3.C1&2). The high growth rate of the subgenus *Passiflora* is a function of a high speciation rate, but accompanied by the highest rate of extinction in the entire genus (Fig. 3.3.D). Conversely, the gradual slowdown in *Decaloba* is driven by a fall in speciation, as the extinction levels have remained low for all of its lineages younger than 12 MA. Most of the credible models in the posterior distribution contain one or two changes in the diversification rate: on the ancestral *Passiflora* branch and sometime in the early history of *Decaloba* (Fig. 3.3.E).

Only 20 species of passion vines have developed a morphological or physiological defence against *Heliconius* besides glycocyanide toxicity (Benson et al. 1975; MacDougal

and Ulmer 2004), including 10 in the supersections *Passiflora* and *Laurifolia* (Appendix). Egg mimics evolved from different morphological parts in nine species (Benson et al. 1975) from six distantly related sections. Trichomes and extreme leaf shape variability have also evolved independently multiple times. Nonetheless, trichomes are the only defence shared by a few closely related species (*P. lobata, P. morifolia, P. adenopoda*).

Figure 3.3. Diversification rate changed between major lineages of *Passiflora.* *A.* Bayesian estimates of net diversification rate of the entire genus *Passiflora* peak around the early diversification of the subgenus *Decaloba,* and has remained very high in the subgenus *Passiflora*. *B.* Lineages Through Time plots show an exponential growth in the number of species. *C.* Diversification rate of *Decaloba* (1) peaked around 10 MA and has decreased sharply since the appearance of the subgenus *Passiflora* 7 MA (2), producing a fairly constant overall rate (3). *D*. Extinction rate has also been high (red) in the subgenus *Passiflora* and old lineages of *Decaloba. E*. Nine top models in the Bayesian credibility set include shifts in diversification rate, typically close to the root of the two major radiations (% frequency reported).


A
net diversification rate
0.6
0.3
0.06
Passiflora (2)
Decaloba (1)
B
All species
Decaloba clade
Passiflora clade
log(N species)
1
2
5
10
20
50
100
200
15
10
5
0
time before present (MA)
C
MA
1
2
3
D
0.5
0.06
0.05
E
6.7%
6.6%
6.3%
4.9%
4.4%
3.1%
2.9%
2.5%
2.1%

Patterns among passion vines

There are 160 species of *Passiflora* attacked by *Heliconius* in the wild. The number of butterflies preying on a passion vine reaches up to 17 (Fig. 3.2). The likelihood of being a *Heliconius* host is not strongly predicted by the passion vine phylogeny: the mean number of predators among the 22 subclades is not significantly different (Kruskal-Wallis rank sum test, X^2=21.94, p=0.145). However, there is weak evidence of higher predation on *Passiflora* (median 5.0 butterflies attacking) than on *Decaloba* (median 2.4; X^2=7.22, p=0.027). A few unrelated species of both subgenera are especially popular hosts, and all are widely distributed and locally abundant (Encyclopedia of Life 2015): *P. laurifolia, P. vitifolia, P. edulis, P. auriculata, P. capsularis* and *P. biflora*.

Host plant usage by Heliconiini

Specific patterns of host plant choice are not obvious. An MDS plotted from feeding preferences does not cluster Heliconiini by clade, but shows that species with a wide geographic distribution tend to have very different diets, which reflect neither their phylogeny nor their range (Fig. 3.4). For instance, allopatric populations of *H. melpomene* cluster neither with conspecifics nor with sympatric taxa. Across Heliconiini the variability is even greater and driven by the wide diets of the older genera and the most widespread *Heliconius* (Fig. S3.3). Little evidence is found for a clade-to-clade mapping, whereby the *H. wallacei* clade supposedly prefers *Distephana*, *melpomene*/Silvaniforms – *Granadilla*, *erato* – *Plectostemma,* and *sapho* – *Astrophea* (Benson et al. 1975; C. Jiggins, *unpublished*). In fact, only 13 among the 39 hosts of *H. melpomene* and relatives are *Granadilla,* the rest coming from the subgenus Decaloba (10 species) and all other clades bar four (Appendix: Associations Table by Clade). Only three out of eight *H. wallacei* hosts are *Distephana*, whereas *H. erato* attacks 17 Plectostemma (Decaloba) hosts, but also 12 *Passiflora* and seven

from other groups. Even the cognates of *H. sapho* attack 10 clades other than *Astrophea*. The number of species in a heliconian clade is obviously not correlated with the total number of hosts attacked (Spearman rank correlation, $p>0.1$; Table 3.4).

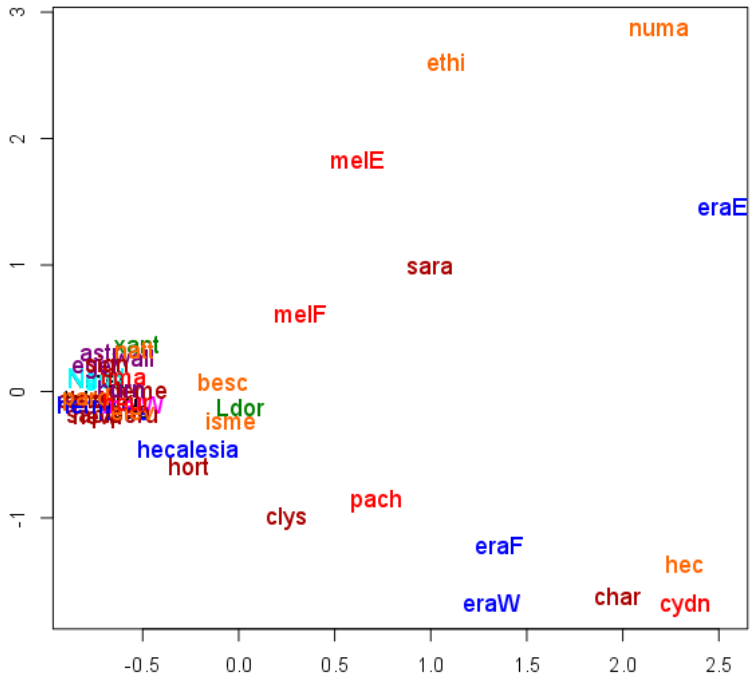


Figure 3.4. Host plant preferences are not similar within *Heliconius* clades. Two-dimensional scaling ordination of Euclidean distances between *Heliconius* based on the host plant usage patterns and colour-coded by clade.

Explicit comparisons of the passion vine and butterfly phylogenies provide statistically significant support for limited codivergence. Parafit indicates family-wide cospeciation at the significance level $p=0.001$. Depending on the significance level, out of 502 associations 246 ($p=0.05$), 138 ($p=0.01$) or 74 ($p=0.001$) are predicted by phylogeny ("due to

codivergence"). For each species I calculate a *Conservatism Ratio* of the number of associations due to cospeciation, compared to the total count of host species. For instance, the three species of *Neruda* all feed on *Dilkea* and thus have a CR of one (*N. aoede: p*=0.01). Conversely, *Eueides procula* feeds on multiple *Decaloba* and *Passiflora,* and none of its preferences are due to cospeciation (CR=0, all $p>0.051$). The ratio has a binomial distribution (Fig. 3.6.B), since few species have a mixture of co-evolved and independently acquired associations. Further evidence for general codivergence is found by the event-based Jane algorithm, which fits an optimal model of Heliconiini evolution on the host tree at the cost of 951, compared to the median cost of 1387.93 for random solutions ($p<0.0001$). The solution consists of 613 host loses, 272 failures to diverge (a butterfly not diverging despite speciation of its host), 28 duplications (butterfly speciation without a host shift), 19 duplications *with* a host switch and only one co-speciation, among the ancestral species of Silvaniforms and the *P. ovalis* lineage. However, the model is too complex to investigate specific events (Fig. S3.4).

The mean number of host plants used differs between related species more than would be expected under a Brownian motion model of trait evolution on the butterfly phylogeny (Blomberg's K=0.43, $p=0.008$). The mean number of hosts is not significantly different between clades (phyANOVA, $p=0.980$; rank sum test, $p=0.100$). Contrary to previous suggestions (Benson et al. 1975), mapping the number of hosts onto the Heliconiini phylogeny reveals no general trend, although the MP ancestral state reconstruction shows generalism in early heliconians (Fig. 3.5.A). Modern sister species can differ tremendously, like *H. hecale* (n=23) and *H. atthis* (n=1).

Clear tendencies are found in the *Conservatism Ratio*, a measure of the diet being predicted by phylogeny, which is demonstrably low among *Philaethria* and *Podotricha*, but fluctuates among *Agraulis*, *Dione* and *Eueides* (Fig. 3.5.B). In *Heliconius* it is consistently

low among melpomene- and silvaniforms, but high in the *H. erato/sapho* clade. Thus the group traditionally described as oligotrophic (Engler-Chaouat and Gilbert 2007) is also conservative in its phylogenetically determined host plant choices. Mean *CR* differs between the main heliconian clades (rank sum test, p=2.195*10^{-06}) (Fig. S3.5).

DISCUSSION

Phylogenetic and statistical approaches showed remarkably little historical signal in the host plant preferences of Heliconiini. Closely related butterflies often feed on very different species of *Passiflora* and *Dilkea* and the majority of heliconian clades contain species feeding on a wide variety of passion vines. This is a surprising finding, as antagonistic coevolution with hosts plants is commonly believed to be a key process generating insect biodiversity (Ehrlich and Raven 1964; Thompson 1999; Fordyce 2010; Occhipinti 2013) and is known to play an important role in the evolution on *Heliconius* and relatives at the level of trait evolution and species coexistence (Benson et al. 1975; Brown 1981; Brower 1997; Merrill et al. 2013). The pattern of phylogenetic correspondence at higher taxonomic levels is well established in lepidoptera (Ferrer-Paris et al. 2013) and a large majority of herbivorous insect families generally appear specialised on a single group of plants (Forister et al. 2015), such as Heliconiini on Passifloraceae. In contrast, modelling shows that only a small fraction of the events during the evolution of heliconian diet were driven by close relations with the passion vines (Fig. S3.4).

The second unexpected finding of this study is the absence of a clear trend in specialisation of the diet among Heliconiini. Position in the phylogeny is not a significant predictor of the number of hosts used by a heliconian species (Fig. 3.5). The ancestor of the tribe was most likely a generalist, although the lability of the observed relations makes it

difficult to speculate what exactly the ancestral diet consisted of. However, I can rule out earlier speculation that the ancestor fed on plants in the subgenus *Distephana*, which is in fact very recent (Fig. 3.2).

Heliconiini are *Passiflora* specialists and it is tempting to speculate that the initial diversification of passion vines, which coincided in time with the divergence of heliconian lineages (Fig. 3.2), may have provided an ecological opportunity for the latter process. However, the comparisons between sister clades of Heliconiini do not support a correlation between speciosity and the diversity of hosts (Table 3.4). Contrary to the results of Fordyce (2010), here and in Chapter 2 I find no clear evidence of an increase in diversification rate of the butterflies driven by host switches. There is also no evidence for the increase in butterfly speciation rate predicted by the oscillation theory of Janz and Wahlberg (2006) (Fig. 2.1), demonstrating that adaptation to plants is not a primary cause of the diversification of Heliconiini. The overall lability in the degree of specialisation is comparable with results from studies of other butterflies (Janz et al. 2006; Janz 2011).

	Smaller clade		**Larger clade**		**Difference**	
smaller clade – larger clade	**B**	**H**	**B**	**H**	**B**	**H**
melpomene – silvaniform	5	39	9	48	4	9
erato – sapho	7	45	11	33	4	-12
Dryas/Dryadula/Podotricha – Philaethria	4	58	7	15	3	-43
Agraulis – Dione	1	37	3	18	2	-19
*Eueides*_1 – *Eueides*_2	3	31	8	50	5	19

Table 3.4 No correspondence between species richness of heliconian clades and the number of *Passiflora* hosts they use. In three cases out of five the larger of sister clades has fewer hosts. B: number of butterflies in the clade. H: number of hosts attacked.

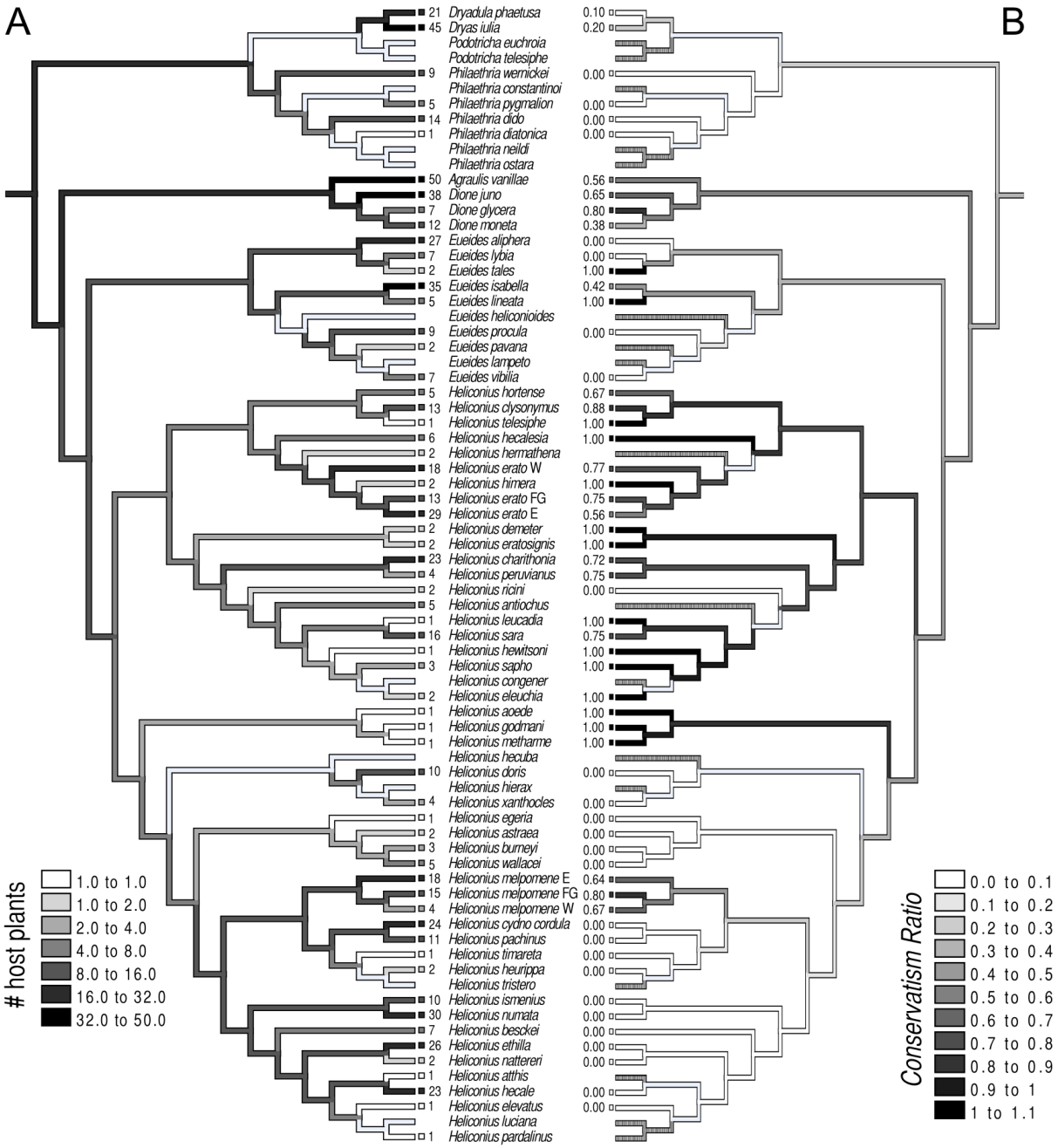


Figure 3.5. The extent of dietary specialisation varies widely in Heliconiini. *A.* The phylogeny of Heliconiini with number of *Passiflora* species attacked (MP reconstruction). No clear pattern was observed. *B.* Mapping of the *Conservatism Ratio* (CR) onto the Heliconiini tree: phylogeny predicts the host plant preference in the *H. erato/sapho* clade.

The only exceptions to the overall instability in the patterns of association are species in the clade of *Heliconius erato* and *Heliconius sapho*. This group, which constitutes around a quarter of the overall diversity in the tribe, tends to have a diet conserved to a much greater extent than any other clade (Fig. 3.5, S3.5). Even though individual lineages can still vary greatly in how many species they take, expansion in the number of hosts is typically not accompanied by shifts to a different clade of *Passiflora*. Variation in the number of hosts is in turn simply correlated to the geographical range of the species (Fig. 3.6), as lineages found over greater areas are more likely to encounter new hosts that they are capable of attacking. The uniqueness of the *H. erato/sapho* group could perhaps be explained by their physiological adaptation to sequester one of the unique classes of *Passiflora* cyanoglycosides – the simple monoglycoside cyclopentenyls (SMCs) (Engler et al. 2000; Engler-Chaouat and Gilbert 2007). Other species of *Heliconius* are unable to perform this feat, but the trade off for the *H. erato/sapho* lineage is their inability to synthesise sufficient amounts of cyanide toxins *de novo* (Engler-Chaouat and Gilbert 2007). These species are therefore largely restricted to a diet that is at least predominantly based on a specific subset of passion vines and provide an excellent example of trade-offs in performance leading to specialisation (Rauscher 1988), which nonetheless does not have to mean a significant reduction in the mean number of species.

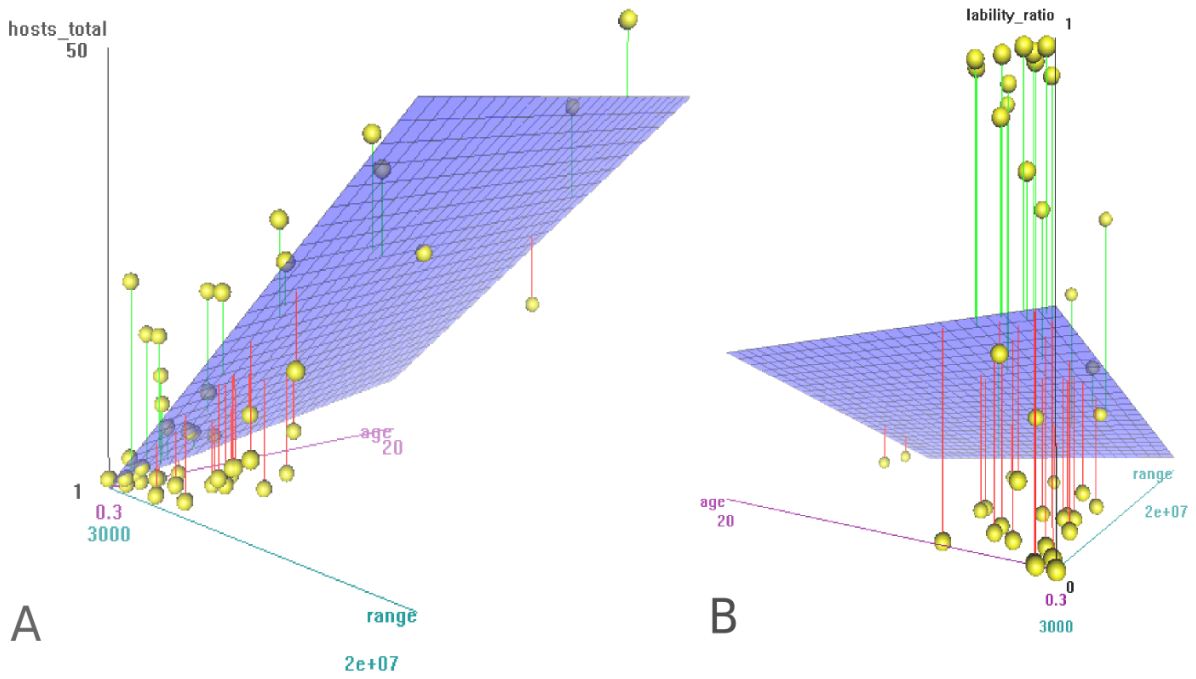


Figure 3.6. Dietary flexibility of Heliconiini correlates with range, but not the time of divergence from the sister species. *A.* Number of plants used by a butterfly against its range area (green axis, km^2) and branch length (violet axis, MA). *B. Conservatism Ratio* against range and branch length (1=diet completely predicted by phylogeny). The *CR* is nearly binomially distributed.

If the preferences are not strictly coevolved outside the *H. erato/sapho* clade, what other factors could influence the larval diet of Heliconiini? Among the more generalist species the ecology influences female oviposition and larval feeding. *Heliconius erato* can be induced to consume non-native host plants without physiological restrictions (Silva et al. 2014). *Heliconius melpomene* can take a wide variety of host species in the Amazon and when raised in captivity, but *H. m. amaryllis* is a specialist in East Peru (C. Mérot, *unpublished*), whereas *H. m. rosina* in Panama feeds only on *P. menispermifolia* to avoid competition in densely

populated habitats (Merrill et al. 2013). Extensive studies of *H. erato phyllis* on the coast of Brazil show that ovipositing females forage optimally by balancing the nutritional value of different *Passiflora* species against their abundance and condition (Rodrigues and Moreira 2002), and this process may contribute to developmental plasticity when variability in larval diets translates into wing shape variance (Jorge et al. 2011). My findings suggest that diet is often a matter of serendipity and a covariate of other ecological factors. It seems unlikely that the wide-ranging species owe their success to their ability to digest a wider range of *Passiflora* [as suggested by (Benson et al. 1975)], since the tremendous variability in range between sister species would therefore imply large changes in physiology at short evolutionary timescales.

Evolution of Passifloraceae

I present the largest phylogeny of Passifloraceae to date and contribute novel insights into the evolution of this charismatic clade of commercial importance. The family Passifloraceae has a Neotropical origin, but various lineages dispersed to the Old World and back to the Neotropics multiple times during the Miocene (Fig. S3.2). Long range invasions were probably facilitated by the pattern of warm oceanic surface currents at the time, which included gyres in the Atlantic and a conveyor belt from South America towards Australasia (Allen and Armstrong 2008). Diversification of the largest genus *Passiflora* bears the hallmarks of divergence driven by environmental change in the Neotropics (Hoorn et al. 2010; Rull 2011). The majority of fossils believed to be Passifloraceae do not meet the criteria for inclusion as calibration points for the molecular clock models (Table 3.1) (Parham et al. 2012), but multiple specimens from the mid-Miocene of Central Europe confirm that at least 16 MA the genus was present in the Old World, outside of its current range, as a minor part of a temperate forest flora (Ševcik et al. 2007; Fossilworks 2015). The fact that a typically

tropical genus is best represented in temperate deposits reflects the poor taphonomic preservation in the tropical environment (Parham et al. 2012). The need to verify the age of fossils carefully is demonstrated by Muschner and colleagues (2012), as well as Hearn (2006) in his work on the Old World *Adenia* (Passifloraceae), who erroneously set the minimum age of *Passiflora* to 37 MA based on a seed fossil (Mai 1967), which was in fact found in Badenian deposits (16.4-13.0 MA) (Ševcik et al. 2007). Some of the lower quality specimens support the presence of Passifloraceae in the Eocene or earlier, which is incompatible with the dates estimated in this study (Fig. 3.2). I suggest that these fossils are either incorrectly identified, like the singular *P. antiqua* specimen from the Cretaceous (Fig. 3.1), or that perhaps the more distant ancestor of Passifloraceae and Malesherbiaceae may have been superficially similar to the former.

I estimate that similarly to Heliconiini (see Chapter 2), the genus *Passiflora* experienced fastest diversification around 10 MA (Fig. 3.3), albeit a the shift in rate cannot be timed precisely. The two major subgenera in the largest genus *Passiflora* have evolved in a strikingly different manner, as the younger subgenus *Passiflora* started to speciate at a very high rate 8 MA later than the more diverse *Decaloba*. Intriguingly, the dramatic rise in the diversification rate of *Passiflora* has been coupled with a decrease in *Decaloba*, pointing to the possibility that the younger clade may be a more effective competitor (Fig. 3.3). Both subgenera show specific adaptations to herbivory by Heliconiini, although it appears that the defence traits are very recent and occur in a relatively small number of species (Appendix: *Passiflora* defences). This supports the contention of Chomicki and Renner (2015) who found surprisingly frequent gains and loses of the protective ant domatia in plants, concluding that the importance of insect-induced morphology on plant diversification may be overestimated. My phylogeny of Passifloraceae could be used to extend the recent work on the evolution of plasticity in the leaf shape (Hearn 2006; Porter-Utley 2014), which appears to be a major

defence against the pattern-searching ability of *Heliconius* (Gilbert 1975) and exemplifies apparent competition mediated by herbivory (Schluter 2000).

Similar to most comparative studies, the presented work is limited by phylogenetic uncertainty and incomplete data. Extensive experiments with the partitioning schemes and modelling found that changes to the number of taxa alone have a pronounced effect on the estimates of divergence times. The dating results in my study are compatible, at least when the 95% HPDs are compared, and have a limited impact on the final conclusions given the similarity of topologies. The hypothesis based on a larger sample of taxa is preferred as less prone to systematic errors (Thomas et al. 2013). Similar uncertainties have been observed before (Sauquet et al. 2012), although the influence of sampling error on molecular clock models is a relatively understudied area that clearly warrants more attention (Filipski et al. 2014). More importantly, the study would benefit from a more extensive sampling of Passifloraceae, as 61/160 taxa attacked by Heliconiini are currently missing from the data. The missing taxa are spread across the clades and constitute only 22.9% of all known associations, making them an unlikely source of large errors.

Through a systematic review of the vast accumulated knowledge on Heliconiini-Passifloraceae interactions, I formally tested many of the pre-phylogenetic verbal hypotheses on the evolution of heliconian host plant usage, while highlighting the complexities that obscure many general trends. Whereas studies of individual divergence events in *Heliconius* variably point to the importance of host plants (Merrill et al. 2013) or lack thereof (Jiggins et al. 1997b), my comprehensive analysis highlights that the signal of coevolution is generally outweighed by frequent host shifts, similar to the obscured coevolutionary state between bees and the flowers they pollinate (Shimizu et al. 2014). Nevertheless, *Heliconius* still provide some of the best examples of coevolution such as egg mimics and hooked trichomes that have arisen in some *Passiflora* (Appendix: Defences) and been overcome by one species of

Heliconius (Cardoso 2008). As is the case for bees, a lack of phylogenetic associations does not rule out coevolution at a trait level.

Although coevolution is an appealing explanation for many ecological interactions, insects may simply colonise the most suitable plants available (Janzen 1980) and specialisation may be a local property of some races of an otherwise generalist species (Fox and Morrow 1981). The flexibility of the heliconian diet suggests that faster-shifting ecological factors often outweigh strict biochemical adaptation when butterflies choose their hosts (Smiley 1978). As pointed out in a recent review, the evidence for coevolution requires both the evidence of a pattern and an understanding of the process (Althoff et al. 2014). By dissecting the pattern I set the stage for future work on the process of passion butterfly adaptation to passion vines, while demonstrating that coevolution played only a small part in generating the staggering diversity of their interactions.

WHOLE GENOME DATA PROVIDE NO EVIDENCE FOR HYBRID ORIGINS OF *HELICONIUS HERMATHENA*

"The great tragedy of Science – the slaying of a beautiful hypothesis by an ugly fact."

- Thomas H. Huxley

An important goal of evolutionary biology is to elucidate how new species arise and get established, a topic studied since the inception of the field but still rife with controversy (Coyne and Orr 2004; Nosil 2012; Seehausen et al. 2014). Recent work has increasingly emphasised the importance of ecological factors in speciation, departing from the traditional model of speciation by geographic isolation (allopatry) and instead highlighting the importance of selection on adaptive traits as the driver of divergence in sympatry and parapatry (Nosil 2012; Feder et al. 2013). This trend has been strengthened by technological advances in genomics, which facilitate identification and quantification of the factors involved in reproductive isolation (Seehausen et al. 2014). A specific type of ecological speciation that remains especially hotly contested in animals is homoploid hybrid speciation (HHS), whereby a new species is formed by an introgression between two parental lineages (Baack and Rieseberg 2007; Schumer et al. 2014). Jiggins and colleagues (2008) propose a narrower category of hybrid trait speciation (HTS), whereby ecological isolation follows from an introgression of genes that both play an adaptive role and confer some reproductive isolation. This mode is contrasted with mosaic genome hybridisation, in which the novel lineage contains loci intrinsically incompatible with both progenitors (Jiggins et al. 2008).

Genomics has made it substantially easier to demonstrate and quantify on-going and historical gene flow across the species barrier. Studies of animal hybrid speciation have proliferated (Abbott et al. 2013) and various degrees of admixture have been found in animal groups as biologically diverse as nematodes (Lunt et al. 2014), fish (Schumer et al. 2013) and great apes (Green et al. 2010) (see Chapter 1). Yet the onus on researchers is now to identify the ecological role of traits controlled by the introgressed variation, and to demonstrate their role in speciation – a burden of proof so far met in only a few systems (Jiggins et al. 2008; Schumer et al. 2014).Genetic studies of butterflies have yielded the most convincing cases (Jiggins et al. 2008; Schumer et al. 2014), owing to the combination of increasingly well-developed genomic resources with deep knowledge of the multifaceted role that prominent wing patterns play in lepidopteran survival and reproduction. Studies of *Papilio appalachiensis*, eponymous with its specialised mountain habitat, revealed a hybrid composition of the genome with ¾ of the variation retained from the cold-adapted *P. canadiensis* (Cong et al. 2015), but *W*-linked Batesian mimicry loci derived from *P. glaucus* (Kunte et al. 2011; Zhang et al. 2013b). The alpine transgressive hybrid of *Lycaeides idas* and *L. melissa* in the Rocky Mountains also shows high altitude adaptation and complex variation in a suite of reproductive and ecological traits (Gompert et al. 2013, 2014).

The genus *Heliconius* is unique in this context, as hybrid trait speciation has been demonstrated in multiple cases across this recent radiation. The adaptive wing phenotype in question contributes to Müllerian mimicry in geographically complex mimicry rings (Jiggins 2008) and may consist of multiple independent elements, controlled by unlinked loci (Sheppard et al. 1985; Counterman et al. 2010; Huber et al. 2015). Not only is there genome-wide gene flow between some species (Martin et al. 2013; Nadeau et al. 2013), but also specific patterning loci under selection introgress between taxa at various levels of divergence (Heliconius Genome Consortium 2012; Pardo-Díaz et al. 2012). The patterning genes are also

linked with mate preference loci (Kronforst et al. 2007; Merrill et al. 2011) and hybrid individuals demonstrate mate preferences for intermediate phenotypes (Melo et al. 2009). Thus wing patterns play a key role in speciation of *Heliconius* (Jiggins 2008) and multiple species fulfil the criteria for hybrid trait speciation (Jiggins et al. 2008; Schumer et al. 2014). *Heliconius heurippa* in particular has been clearly documented as a hybrid species founded through an adaptive introgression of the red pattern alleles from *H. melpomene* into the *H. cydno* background (Salazar et al. 2008, 2010), to form a unique non-mimetic phenotype established in geographic isolation and mating assortatively (Mavárez et al. 2006).

Based on the phenotype it has been proposed that the rare *H. hermathena* originated through homoploid hybrid speciation (Jiggins et al. 2008), which so far has not been documented among the 20 species in the *H. erato* clade, the largest among *Heliconius.* To date, all cases of adaptive introgression have been described in the more recently diverged *H. melpomene/cydno/*Silvaniform clade. The notable wing pattern of *Heliconius hermathena* comprises fine longitudinal yellow bars ("zebra"), most commonly seen in *H. charithonia*, combined with the broad red forewing band known from several species across the radiation, especially the races of *H. erato* outside the Amazon basin (Fig. 4.1) (Brown and Benson 1977; Jiggins et al. 2008). It is also an ecologically peculiar, non-mimetic species found only in the scrub (*pseudo-caatinga*) and savannah (*cerrado*) habitats on sandy soils along the Amazon catchment rivers (Fig. 4.1), and the only herbivore exploiting the woody *Passiflora hexagonocarpa* and *P. farroana* (Brown and Benson 1977). Unlike most Heliconius, which are mimetic throughout their geographic range and spectrum of phenotypic variation, five out of six *H. hermathena* races have subtle variations of the non-mimetic transgressive pattern. Only at one location *H. hermathena verreata* develops almost no yellow markings and resembles *H. melpomene melpomene* and *H. erato hydara* (Fig. 4.1), which it encounters at the rainforest-grassland ecotone (Brown and Benson 1977).

H. charithonia

H. hermathena

H. h. verreata

H. erato

H. himera

Figure 4.1. **Distribution of *H. hermathena*** and the putative parental taxa based on museum records (Rosser et al. 2012). Red forewing banded races of *H. erato* shown: rayed forms occupy the gap in the Amazon basin. Topographic map generated at www.cartodb.com (2015) with a Nokia Day© background.

Heliconius hermathena was first described by William Chapman Hewitson (Hewitson 1854), an accomplished collector responsible for several descriptions of *Heliconius* and co-mimics (Fig. 2). “Illustrations of new species of exotic lepidoptera” (p. 150) discloses nothing about the inspiration behind the name, but modern authors have speculated that it may reflect the seemingly hybrid pattern through an allusion to *Hermathenae* (Beltrán and Brower 2010, Tree of Life Web Project 2015) - once widespread effigies of the divine Hermes and Athena, united to symbolise duality (Fig. 4.2) (Kelly 2009). *Heliconius* have been named after mythical figures since Linnaeus (1758) and Kluk (1780), and a classically educated Victorian gentleman would surely be aware of the references, his fondness for the name further reflected in naming an unrelated Riodinid genus *Hermathena* (1874), despite its entirely different appearance.

Whatever the provenance of the suggestive name, introgression is plausible, since one apparent *H. erato* x *charithonia* hybrid is known from museum collections (Mallet et al. 2007). *Heliconius himera*, a recent and frequently hybridising derivative of *H. erato* (Jiggins et al. 1997a; Mallet et al. 2007), was observed mating with *H. charithonia* at the insectaries in Gamboa, Panama in 2015, although the female did not survive to lay eggs (C. Jiggins, *pers. obs.*). Hybridisation may be a rare occurrence at ~6 MA of divergence between the putative parental species (Chapter 2), and crossing alone cannot demonstrate introgression of genetic variation beyond a transient presence (Baack and Rieseberg 2007).

Figure 4.2. Hermathena as the quintessential hybrid. Left: the original watercolour of *Heliconius hermathena (top), H. hecalesia* and *H. heurippa* by Charles Standish, accompanying he descriptions by William C. Hewitson (1854). It is probably a coincidence that all three species are suspected hybrids. Right: deities conjoined into a Hermathena, as depicted in the dialectic treatise by Mechovius (1677).

Multiple authors have argued that morphology serves as a valuable indicator of hybridisation, as the product of multiple genetic factors at once (McDade 2000; Arnold 2004). Indeed, wing patterns triggered the discovery of adaptive introgression in *H. heurippa* (Mavárez et al. 2006; Salazar et al. 2010) and *H. elevatus* (Heliconius Genome Consortium 2012; Wallbank et al. 2015). Brown (1981) classifies *H. hermathena* as more closely related to *H. charithonia* than *H. erato* based on unspecified external characteristics, even though its male genital traits and the distribution of androconia are more similar to *H. erato*, whereas the lack of signa on the female bursa copulatrix is a synapomorphy of the entire subgenus (Brown 1981; Holzinger and Holzinger 2000). Caterpillars and pupae of the species have unique

morphologies, including striking spots and bright red larval coloration, which (Brown and Benson 1977) may simply be derived adaptations to the unusual grassland habitat. The hypothesis of hybrid speciation from *H. charithonia* x *erato* crosses has been entertained recently on the basis of discordant signals between single nuclear and mitochondrial data (Beltrán et al. 2007), but more sequence data is needed for verification. My analysis in Chapter 2 does not support admixture with *H. charithonia*, but relies on a small fraction of the genome, specifically not including any of the patterning loci.

Explanation of *H. hermathena* speciation is further relevant for dissecting the history of the largest mimicry complex in *Heliconius*, which involves the repeated convergence between *H. erato* and *H. melpomene*. The two radiations occupy almost the entire tropical range of the genus (Fig. 4.1) and comprise ~30 distinct wing pattern races each. Early work focused on the possibility of divergence in temporary allopatry in forest refugia during Pleistocene climatic oscillations (Brower 1996), and later work has revisited the importance of changing geophysical features such as the rising Isthmus of Panama (Quek et al. 2010; Hill et al. 2013). Subsequent studies showed that although neutral markers can be informative about demography (Flanagan et al. 2004) and dispersal (Quek et al. 2010), selection causes the patterning loci to spread independently. Thus the colour pattern race divisions do not necessarily illustrate genome-wide divergence (Hines et al. 2011; Supple et al. 2013; Nadeau et al. 2015). It remains contested whether the two species evolved their mimetic races in parallel (Brower 1996; Cuthill and Charleston 2012), or if perhaps *H. melpomene* has adverged on more ancient variation in *H. erato*. The work of Quek and colleagues (2010) purports to provide evidence for the latter by showing that the two species diffused across South America from the opposite directions, but the argument hinges on poorly supported AFLP phylogenies. However, Supple et al. (2013) demonstrate a single origin of the Amazonian *Dennis/Ray* alleles in *H. erato*, supporting late evolution of these morphologies

upon contact with older rayed Heliconiini (e.g. *Eueides procula, Heliconius aoede, H. demeter*). As a close relative sharing one of the most common patterns, *H. hermathena* may provide key evidence for the origins of *H. erato*.

In this chapter I test the hypotheses regarding the genomic composition of *H. hermathena* with data from the exome and non-coding loci, assessing (i) systematic affinity of the species; (ii) evidence for genomic mosaicism; (iii) signatures of adaptive introgression at the major wing pattern loci. Admixture is distinguished from other population-level processes through a combination of coalescent approaches. For the first time I provide a genome-wide, statistically robust analysis of biogeography and mimetic history of the *H. erato* complex, with an emphasis on the understudied yellow/white patterns.

METHODS

Genome resequencing and genotyping

Sampling: Novel data were generated for 44 individuals in 14 species, including three samples of the putative hybrid *H. hermathena*, 10 allopatric races of *H. erato* and two *H. charithonia*. In addition, I used 35 libraries from previous studies (Heliconius Genome Consortium 2012; Briscoe et al. 2013; Martin et al. 2013; Supple et al. 2013; Wallbank et al. 2015). The final collection comprises 72 specimens from 18 species in the *H. erato* clade, and seven samples of four outgroup species at various levels of divergence (Supplementary Table 1). I also included 13 races of *H. erato* from across its range, sampling multiple individuals per population where possible (Appendix).

Sequencing: Novel libraries were prepared in the laboratories of W. O. McMillan (STRI) and J. Mallet (UCL) according to the protocol reported by Supple et al. (2013). 100 bp paired-end reads were sequenced with the Illumina Genome Analyzer II and HiSeq2000 platforms with the average insert sizes between 200 and 400 bp. Sequence quality was examined individually with FastQC v. 0.8 (Andrews 2014) and the portions of the reads with base quality scores Q<20 were trimmed using fastx_trimmer v. 0.13 from the FASTX-Toolkit (Gordon 2009).

Mapping: There is currently no whole-genome reference for *H. erato.* I used the genome of *Heliconius melpomene* (Heliconius Genome Consortium 2012), which split off around 12 MA (Chapter 2) and evolved approximately 10-13% genome-wide divergence (S. Martin, *pers. comm.*). The Illumina paired-end reads were mapped using the *aln* and *sampe* algorithms in BWA v. 5 (Li and Durbin 2009), which are sensitive enough to enable the mapping to the conserved parts of the genome, such as the exome (Davey 2013; Supple et al. 2013). Based on preliminary runs, I reduced the minimum seed length parameter *k* from 32 to 25 to facilitate mapping to the divergent reference.

Genotyping: The .bam files were sorted and indexed in Samtools (Li et al. 2009). Picard Tools were used to fix the read group headers, mark optical duplicates and fix the read pair information (Fennell 2010). Local alignment of indels was performed individually with the Genome Analysis ToolKit (GATK) v. 2.6, following the standard guidelines (McKenna et al. 2010; van der Auwera et al. 2013). The Base Quality Score Recalibration step (DePristo et al. 2011) was omitted because no databases of reference variants were available at the time. However, since my conclusions are based on genome-wide phylogenetic analyses, the relative quality of individual SNPs is not a major concern. Read mapping statistics were calculated using the Samtools *flagstat* command. The full bash script with complete commands can be found in Supplementary Data.

As a compromise between increased accuracy and sample-specific biases, all individuals of one species were genotyped together with the UnifiedGenotyper algorithm in GATK v. 2.6 (*-emit_all_confident sites,* default filters). Species VCFs were joined with *vcftools merge*. Using custom python scripts provided by Simon Martin, I converted the VCF into the tabular calls format (Martin et al. 2013), filtered out sites with coverage <10x and quality Q<30, eliminated columns with data missing for more than 19/39 individuals (supermatrix **SM50**) or any missing data (supermatrix **SM100**), and generated genome-wide fasta alignments.

Whole genome phylogeny

Variation in the supermatrices was estimated in PAUP* v. 4 (command *cstat*) (Swofford 2002). Maximum Likelihood phylogenies were generated in RAxML v. 7.2.8 under the GTR+GAMMA model with 100 bootstrap replicates (Felsenstein 1985; Stamatakis 2006). The program was executed on the *Butterfly* server in the School of Life Sciences, University of Cambridge, using SSE3 hyperthreading on 20 AMD-Opteron 6380 CPUs. The mitochondrial data were analysed separately.

CDS gene trees

I reconstructed the species tree and detected hybrids based on gene trees estimated from individual CDS alignments. To reduce computational intensity I limited the dataset to 38 individuals by excluding redundant samples from the same population, and the poorly mapped *Eueides tales*. GATK genotyping was repeated with the *-emit_all_sites* option and separate alignments of CDS gene annotated in the reference (Heliconius Genome Consortium 2012) were extracted with python scripts (Briscoe et al. 2013). Alignment sites were filtered individually (depth≥10x, Q≥30, missing data <50%). Illumina reads from recently diverged

paralogues may potentially map to the wrong reference loci, resulting in hybrid alignments and erroneous inference of gene trees (Teo et al. 2012; Nadeau et al. 2013). I narrowed the gene set to the loci shown to be 1:1 orthologs between *H. melpomene*, monarch butterfly *Danaus plexippus* (Zhan et al. 2011), silkworm *Bombyx mori* (Xia et al. 2004) and fruit fly *Drosophila melanogaster* (Pruitt et al. 2005), and thus unlikely to be candidates for recent duplication in *Heliconius* (Heliconius Genome Consortium 2012). 7236 alignments were left after the exclusion of the 87 sex-linked *Z* chromosome genes (Martin et al. 2013) and genes on the four large scaffolds containing the *B*/*D*, *Yb*/*Cr* and *Ac*/*Sd* wing pattern loci were also removed (Supplementary Table S2).

TrimAl v.1.2 was used to remove sites with more than 50% missing data from individual alignments (Capella-Gutiérrez et al. 2009). As many of the genes were recovered only in small portions, I focused on a core set of 1000 longest alignments (**E1000**). Studies with simulated and empirical data demonstrate that the resolution of gene trees correlates with the length of the alignment more than the substitution rate (Aguileta et al. 2014; Thiergart et al. 2014) and that a few hundred loci, including a small number of highly-variable genes, is sufficient to disentangle many recent radiations (Lemmon and Lemmon 2012; Lanier et al. 2014). Individual gene trees were estimated in PhyML v.3.1 (Guindon et al. 2010) under the GTR+G model of substitution using SPR moves. To minimise the error introduced by uninformative alignments and incorrectly estimated trees, I used Mesquite v. 2.75 (Maddison and Maddison 2006) to filter out trees where the ingroup was polyphyletic, reducing the data to 921 trees.

Multispecies coalescent phylogeny

Gene flow between species, as well as gene loss and duplication, heterogeneity of branch lengths and incomplete lineage sorting, may all mislead the inference of a phylogenetic tree under the simple assumption that all parts of the genome share one history and can be joined into a supermatrix (Chapter 2; Edwards 2009; Nakhleh 2013). Multiple methods have been proposed to account for this problem under the multispecies coalescent framework (Knowles and Kubatko 2010; Yang and Warnow 2011; Leaché et al. 2013). To establish a general species tree fast, I integrated over the 921 gene trees using the analytical Minimising Deep Coalescence (MDC) method (Maddison 1997) in the package PhyloNet v.3.4 (Than et al. 2008). Phylogenies were rooted by the outgroups, branch lengths were included in the calculations, samples were assigned to species ("*-a taxa map*") and 100 bootstrap replicates were executed.

Tests of introgression

The ABBA/BABA test is used to detect the signal of ancient admixture across a genome by testing for significant imbalance in the frequencies of incongruent gene trees (Green et al. 2010). In a four-taxon case with an outgroup (e.g. *H. melpomene*), two sister taxa (P_1 and P_2; *H. erato* and *H. hermathena*) and a third ingroup species (P_3; *H. charithonia*), we expect most sites to show the AABB pattern, where A is the ancestral state and B is the derived state shared by the most recently diverged species P_1 and P_2. Due to incomplete lineage sorting some sites may also show a deviant pattern (ABBA or BABA), but these stochastically occurring alternatives are expected at equal frequency. If gene flow took place between the distantly related taxa (P_2 and P_3), we expect one of these topologies to occur in excess. The difference can be quantified in terms of the D and *f* statistics (Durand et al. 2011), although their properties and robustness to violations of the underlying assumptions are only

beginning to be explored (Martin et al. 2015). As *H. erato* and *H. hermathena* are sister species (see below), the most interesting scenario is gene flow to and from *H. charithonia*, leading to the introgression of yellow wing pattern loci into *H. hermathena.* I treated *H. charithonia* as the donor of gene flow (P_3), *H. hermathena* as the recipient (P_2), *H. erato* as the non-recipient (P_1) and *H. melpomene* as the outgroup. To account for the population structure in *H. erato*, I also conducted separate tests with Amazonian, Guianian (sympatric with *H. hermathena*) and West of Andes (allopatric) populations of this species. Gene flow between the putative progenitors was tested with *H. sara* (P_1), *H. charithonia* (P_2) and *H. erato* (P_3). Robustness of the technique to Type I error was tested by substituting the donor (P_3) with *H. sara* or *H. clysonymus*, which are not suspected contributors to the *H. hermathena* genome.

All the ABBA/BABA tests were conducted on the genome-wide **SM50** supermatrix with all individuals of each species, using the python and R scripts by Martin and colleagues (2013) to analyse the derived allele frequency at biallelic sites. Tests were conducted separately on the 20 autosomes, the sex-linked Z chromosome and the 18 scaffolds linked to colour pattern loci in *H. erato* (Nadeau et al. 2015). Significance of the test statistics in the face of linkage disequilibrium (LD) was assessed using a block jacknife procedure (Reich et al. 2009) with the window size of 1 Mbp (100 kbp at the pattern loci), substantially exceeding the known amount of LD in *Heliconius* at the relevant level of divergence (Heliconius Genome Consortium 2012; Martin et al. 2013).

The hypothesis of hybridisation was also tested under the coalescent model of Yu et al. (2011, 2013), which distinguishes between the signatures of gene flow and unsorted ancestral variation under the Maximum Parsimony criterion. The algorithm implemented in PhyloNet v. 3.4 (option "*-InferNetwork_parsimony*") was applied to the **E1000** dataset. Up to ten equally parsimonious reticulation networks were inferred, and visualised Dendroscope v. 3

(Huson and Scornavacca 2012).

A complementary way to test for introgression and its direction is to compute the likelihood of individual gene trees deviating from the species tree, accounting for incomplete lineage sorting. Incongruent phylogenies grouping *H. hermathena* and *H. charithonia* together were detected with Riata's Horizontal Gene Transfer test (HGT) (Nakhleh et al. 2005) in PhyloNet.

Analysis of colour pattern loci

I used the *H. erato* bacterial artificial chromosome (BAC) sequences reported by Papa et al. (2008) as the reference for read mapping to the *B/D* and *Cr* patterning loci. BACs for the unlinked downstream effector *Cinnabar* (Ferguson and Jiggins 2009) and clones 48A16, 46F09 were included to reduce the number of reads from repetitive elements. Reads from the 38 focal individuals were mapped with BWA v.5 with default parameters, except for the relaxed seed length parameter *k*=25 for the outgroup species. Confident sites were genotyped in GATK as above.

The *Cr* locus is currently represented by three non-overlapping BAC clones. An alignment to the *H. melpomene* sequences suggests that there are long gaps between the BAC sequences and the peak of diversity (F_{st}) among differently patterned races is located in one of the gaps (Nadeau et al. 2012; Supple et al. 2013). I attempted to fill the gaps by iteratively joining *de novo* contigs from Chapter 2 to BAC sequences in Geneious v. 5 (Biomatters Ltd 2012), but at most 30% of the expected sequence was recovered (results not shown). The three *H. erato* BACs were instead aligned against the continuous *H. melpomene* sequence using mLagan (Brudno et al. 2003) and visualised in VISTA (Mayor et al. 2000). *Heliconius melpomene* sequence was inserted into the two gaps of 174,387 bp and 47,446 bp to create a hybrid *Cr/Yb* reference. In the absence of a complete *Sd* locus reference, I also

mapped to the BAC for the homologous *H. melpomene Ac* locus.

Studies investigating very closely related species often compare phylogenies in equally sized sliding windows (e.g. Martin et al. 2013). Mapping divergent taxa like *H. charithonia* to the BAC references results in variable coverage and tracts of missing data, making an optimal partitioning approach more appropriate (Supple et al. 2013). I applied the parsimony-based minimum description length algorithm (MDL), which detects recombination events and is computationally tractable for many taxa (Ané 2011). I set the minimum size of a partition to 10 kbp, based on the LD in the *H. melpomene/cydno* clade and the typical size of critical colour pattern loci ranging from 5-10 kbp (Nadeau et al. 2014; Wallbank et al. 2015). ML trees were estimated in RAxML v.7.2.8. If introgression was suspected, I conducted the SH test (Shimodaira and Hasegawa 1989) to compare against the species tree. Hybridisation was distinguished from ILS using the ABBA/BABA test in 20 kbp windows sliding by 10 kbp, with *H. sara* as the outgroup to minimise missing data. Additionally, coding and non-coding sequences of the key yellow patterning gene *poik* (Nadeau et al. 2015) were used to build ML trees and conduct SH tests of the hypotheses of *H. charithonia* alleles as: sister to *H. hermathena* (following introgression), basal to the *H. erato* or *H. sara* clades, or in the species tree position. Selection on the CDS was tested in HyPhy (Pond et al. 2005).

RACIATION OF *H. ERATO*

To analyse the patterns of divergence within the large radiation of *H. erato* subspecies, I estimated an ML phylogeny as above for the long supermatrix (sites with <50% missing data) comprising 13 wing pattern races from 18 populations ranging from Mexico to Bolivia. Times of divergence were estimated under the relaxed molecular clock (UCLD) model in BEAST v. 1.8 (Drummond et al. 2006, 2012) on the fixed ML topology. Using results from Wahlberg et al. (2009) and Chapter 2, I specified the following uniform priors on split ages:

Heliconius/Eueides (16.5-20.6 MA), *H. melpomene/H. erato* (10.5-13.4 MA), *H. erato/H. sapho* (5.4-7.1 MA). I ran three independent chains of $15*10^6$ cycles of MCMC with 20% burnin, assessing the convergence and ESS of continuous parameters in Tracer v. 1.6 (Rambaut et al. 2014). Due to computational constraints, the dating was based on 10^5 sites sampled from the original alignment.

Since gene flow is expected, a standard bifurcating tree may not be a sufficient representation of the relations between populations (Huson and Bryant 2006). I used networks to investigate the amount of structure in the *H. erato* data. Pairwise genetic distances were calculated in SplitsTree v. 4 (Huson and Bryant 2006) under the most complex tractable model (F84) to infer a split network (Steel 2005). A consensus of 1000 bootstrap replicates was taken to assess confidence.

The loci known to regulate the adaptive patterning have a different history from the putatively neutral background across *H. erato* (Hines et al. 2011), so I generated separate ML phylogenies for the critical intervals: positions 300,000-450,000 on the *D* scaffold (Supple et al. 2013; Wallbank et al. 2015) and 670,000-971,000 on the *Cr* scaffold (Nadeau et al. 2015). As whole genome data were available only for a limited number of populations, a phylogeny of 374 GenBank samples and 149 sequences from Chapter 5 (partially overlapping with this work) was estimated for *Optix*, a key gene at the *D* locus (Reed et al. 2011; Pardo-Díaz et al. 2012), after an alignment in MUSCLE (Edgar 2004). To better understand the history of the *Cr* locus, I fitted an AMOVA model (Excoffier et al. 1992) in the R package *pegas* (Paradis 2010) with pattern, race and geography as predictors.

RESULTS

Sequencing and mapping

Illumina data was collated for 79 individuals in four outgroups and 13 species of the *H. erato* clade, including 13 wing pattern races of *H. erato*. Samples were available from one population of *H. hermathena hermathena* in the vicinity of Santarem. In total 47.86% of the high quality (Q≥30) genome-wide data came from the exome and 52.14% from the non-coding regions. There were 48,502,935 SNPs (5,644,783 multiallelic), but after filtering out sites with mostly missing data the supermatrices SM50 (17,175,625 bp) and SM100 (943,660 bp) contained only 17.17% and 0.09% variable sites (Table 2). Hence the phylogenetic and ABBA/BABA analyses were based on SM50. After quality filtering of sites in the VCF and individual CDS alignments (Q≥30, <50% missing data) there were 5228 alignments longer than 100 bp. I focused on a subset of 1000 longest genes, 921 of which contained the monophyletic ingroup (Table 2). Out of 15,327 sites in the mitochondrial alignment 10,842 passed the quality filters. Mapping to the *H. erato* and hybrid *H.erato/melpomene* BAC references for the colour pattern loci was moderately successful. After filtering I recovered 177721 bp at the *D* locus, 119261 at the *Cr* and 14195 at the *Sd*.

	All genes	E1000 (longest)	SM50
# alignments*	5228	921	supermatrix
Length mean (range) bp**	648.26 (100-11236)	1,741.85 (894-11236)	17,175,625
Variable sites	86.55 (0-5532)	266.27 (43-5532)	2,948,330
Parsimony informative	123.37 (0-4690)	123.37 (22-4690)	1,522,168
Pairwise identity	90.5%	91.9%	84.3%
Missing/gap/ambiguous	5.3%	4.0%	11.8%
GC content	38.0%	37.4%	34.6%

Table 4.1. Variability in the single-copy protein-coding gene alignments of 39 taxa and the genome-wide supermatrix of 79 samples. *Alignments with a polyphyletic ingroup excluded. **After trimming off sites with >50% missing data. Alignments under 100 bp excluded.

Heliconius hermathena is the sister species of Heliconius erato

*C*oalescent and supermatrix trees, networks and explicit tests of hybridisation do not support the hypothesis of mosaic genome composition in *Heliconius hermathena*. The MSC tree of 921 exome markers with the MDC algorithm (the species tree) confirms that *H. hermathena* is the sister species of *H. erato/himera* and is only distantly related to *H. charithonia*, which forms part of the large *H. sapho* clade sensu Brown (1981). This result is further supported by the mitochondrial data (Fig. SF4.1) and the concatenated genome-wide matrix (Fig. SF4.4), which uncover relations consistent with those found in Chapter 2. The genomic background of *H. hermathena* is therefore similar to that of *H. erato,* whereas *H. charithonia* is likely to be at best the minor donor of admixed variation.

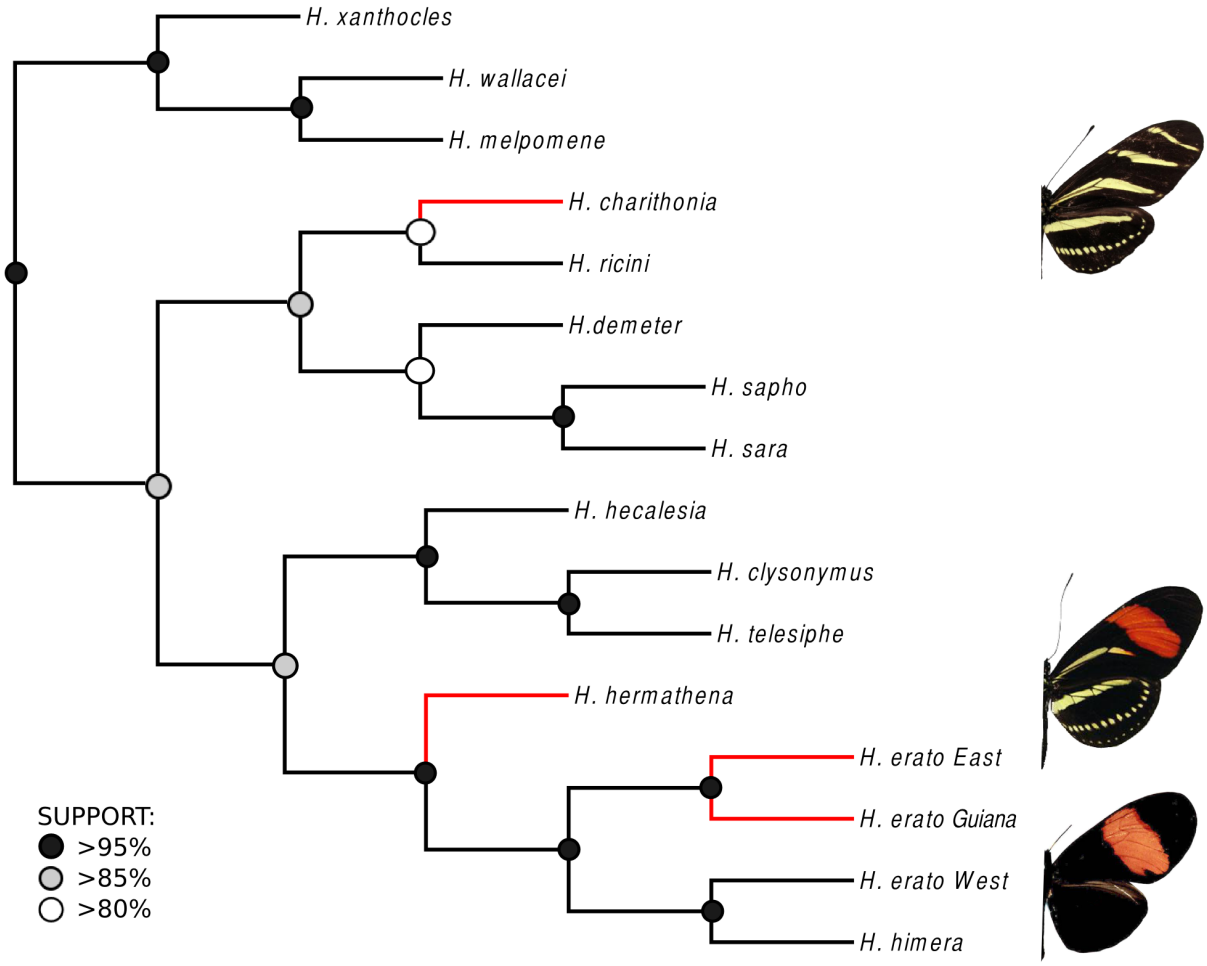


Figure 4.3. *Heliconius hermathena* is the sister species of *H. erato* and distantly related to *H. charithonia*. A coalescent (MDC) tree based on 921 CDS gene trees with bootstrap support values.

No evidence for a Heliconius hermathena mosaic genome

The above analyses may easily fail to provide evidence of hybridisation, as even concatenation of conflicting markers is likely to yield well-supported trees given enough data (Jarvis et al. 2014; Thiergart et al. 2014; Chapters 2 and 5), and most MSC techniques assume little to no gene flow (Nakhleh 2013). Reticulation networks (Yu et al. 2012), designed to address these shortcomings, also place *H. hermathena* away from *H. charithonia* and show no convincing evidence of admixture at loci of no known strongly selected function (Fig. S4.2). Curiously, some evidence is found for *H. erato* x *H. telesiphe* hybridisation producing *H. hecalesia,* and *H. demeter* x *charithonia* yielding *H. ricini* (Fig. S4.2) (see Chapter 5 for an in-depth analysis).

The complementary graph-theoretic test of Riata (Nakhleh et al. 2005) finds introgression from *H. charithonia* at two out of 921 gene trees. Intriguingly, one of these loci, HMEL009230 (scaffold 670877, chromosome 18) encodes a homolog of the zinc-finger transcription factor Rotund, which in *D. melanogaster* regulates the development of imaginal disks by initiating the expression of *Wingless* at the hinge, and is required for eye development (Herrera et al. 2013). HMEL017745 (scaffold 671647, chromosome 16) encodes an ADP-dependent glucokinase of unknown function. It must be noted that Riata's test may be unreliable, as it yields evidence of unexpected admixture between several other species.

Computation of the ABBA/BABA statistics is considered a powerful method for identifying gene flow across the species barrier at the genome level, but in this study yields unexpected and conflicting results. However, my work demonstrates that the test can be hindered by an astounding false discovery rate. The *D* statistic for *H. charithonia* x *hermathena* gene flow at the autosomes is estimated at 0.073 (jacknife error estimate ±0.005, p=8.8*10^{-45}), and the amount of genome-wide admixture f_4 (Patterson et al. 2012) at ~0.010 (±0.001) (Table 4.2.A). However, there is no evidence for the hypothesis of *H. charithonia* x

H. hermathena gene flow when *H. sara* is the non-recipient instead of *H. erato* (*p*=0.17; Table 4.2.B). The biological significance of the ABBA/BABA tests is questionable, given that the highest value of *f* (0.015; D=0.07, p=3.14*10^{-43}) is estimated for the made-up control scenario of gene flow between *H. hermathena* and *H. clysonymus* (Table 4.2.D). Furthermore, despite the expectation of heterogeneity in gene flow across the genome (Martin et al. 2013), the statistics are similar for autosomes, the sex chromosome and specific wing pattern loci across the range of tests (Table S4.1). The only notable exception is the admixture between *H. charithonia* and *H. erato* restricted to the autosomes (*D*=0.019, *p*=9.6*10^{-5}).

	non-recipient (P_1)	*recipient (P_2)*	*donor (P_3)*	Test	*chrom*	*D*	*D err*	*D Z*	*D p*	*f*	*f err*
A	*erato*	*hermathena*	*charithonia*	*H. charithonia* x *H. hermathena* gene flow	Auto	**0.073**	**0.005**	**14.041**	**0.000**	**0.010**	**0.001**
					Z	**0.073**	**0.010**	**7.316**	**0.000**	**0.010**	**0.001**
					Pattern	**0.081**	**0.029**	**2.782**	**0.005**	**0.012**	**0.004**
B	*sara*	*charithonia*	*hermathena*	*H. charithonia* x *H. hermathena* gene flow	A	0.009	0.007	1.368	0.171	0.002	0.001
					Z	0.009	0.018	0.484	0.629	0.002	0.003
					P	-0.033	0.042	-0.796	0.426	-0.006	0.007
C	*charithonia*	*sara*	*hermathena*	Control: unlikely recipient	A	-0.009	0.007	-1.368	0.171	-0.001	0.001
					Z	-0.009	0.019	-0.484	0.629	-0.001	0.003
					P	0.033	0.042	0.796	0.426	0.006	0.007
D	*erato*	*hermathena*	*clysonymus*	Control: unlikely donor distributed like *H. charithonia* (West of Andes)	A	**0.070**	**0.005**	**13.785**	**0.000**	**0.015**	**0.001**
					Z	**0.070**	**0.021**	**3.396**	**0.001**	**0.015**	**0.005**
					P	0.158	0.029	5.370	0.000	0.033	0.006

Table 4.2. ABBA-BABA tests do not provide credible evidence for the introgression hypothesis. Equally strong evidence is found for gene flow between *H. hermathena* and *H. charithonia* as between the implausible *H. hermathena – H. clysonymus* pair. Signal (*D*) and proportion (f_4) of genome-wide gene flow was estimated with a *H. melpomene* outgroup. Values reported separately for autosomal chromosomes (A), the Z and 18 pattern-linked (P) scaffolds. Significant results in bold. Values for all 54 tests reported in Table S4.1.

No evidence for adaptive introgression at Heliconius hermathena wing pattern loci

Taken together, results of Riata's test and *D* statistic estimation suggest that if introgression from *H. charithonia* occurred, its scope was extremely limited. It is possible that preferential backcrossing of F1 hybrids to *H. hermathena* and recombination have purged the admixed variants, except the ones favoured by selection. To test this hypothesis I calculated *D* statistics and inspected gene trees between recombination breakpoints across the adaptive patterning regions *D*, *Cr* and *Sd* (Sheppard et al. 1985). I found only a handful of sites with an excess of ABBA or BABA sites and thus the emerging patterns of the *D* and *f* statistics were predictably chaotic (Fig. S4.3). The observed patterns are similar to those at the *Cinnabar* locus, which is not known to be under strong positive selection.

The MDL algorithm (Ané 2011) found eight to 12 recombination breakpoints in the three pattern loci alignments (Table S4.1). Gene trees across those regions were generally congruent with the species tree, with most deviations unsupported by bootstrap. The placements of *H. hermathena* and its putative parental species are almost perfectly stable at the red pattern locus *D*. *Heliconius charithonia Cr* and *Sd* (yellow/white) alleles are often found at the root of the entire *H. sapho/erato* clade, but the species tree topology is equally likely at the p=0.01 significance level (SH test). The coding *poik* sequences from *H. charithonia* cluster it with *H. hermathena* together, counterintuitively suggesting introgression *from* the latter. However, this relation does not withstand the SH test (p<0.01) and disappears when the 9,511 bp of intronic sequence is included. The clustering may be caused by molecular convergence in the protein-coding sequence, as Tajima's *D* is negative (-2.268) and HyPhy detects two codons under purifying selection (579 bp, p=0.019; 861 bp, p=0.004). The gene tree of the 1100 bp *Optix* CDS, including 513 individuals, also shows that *H. hermathena* alleles are deeply diverged from *H. charithonia* and sister to *H. erato* (Fig. S4.6).

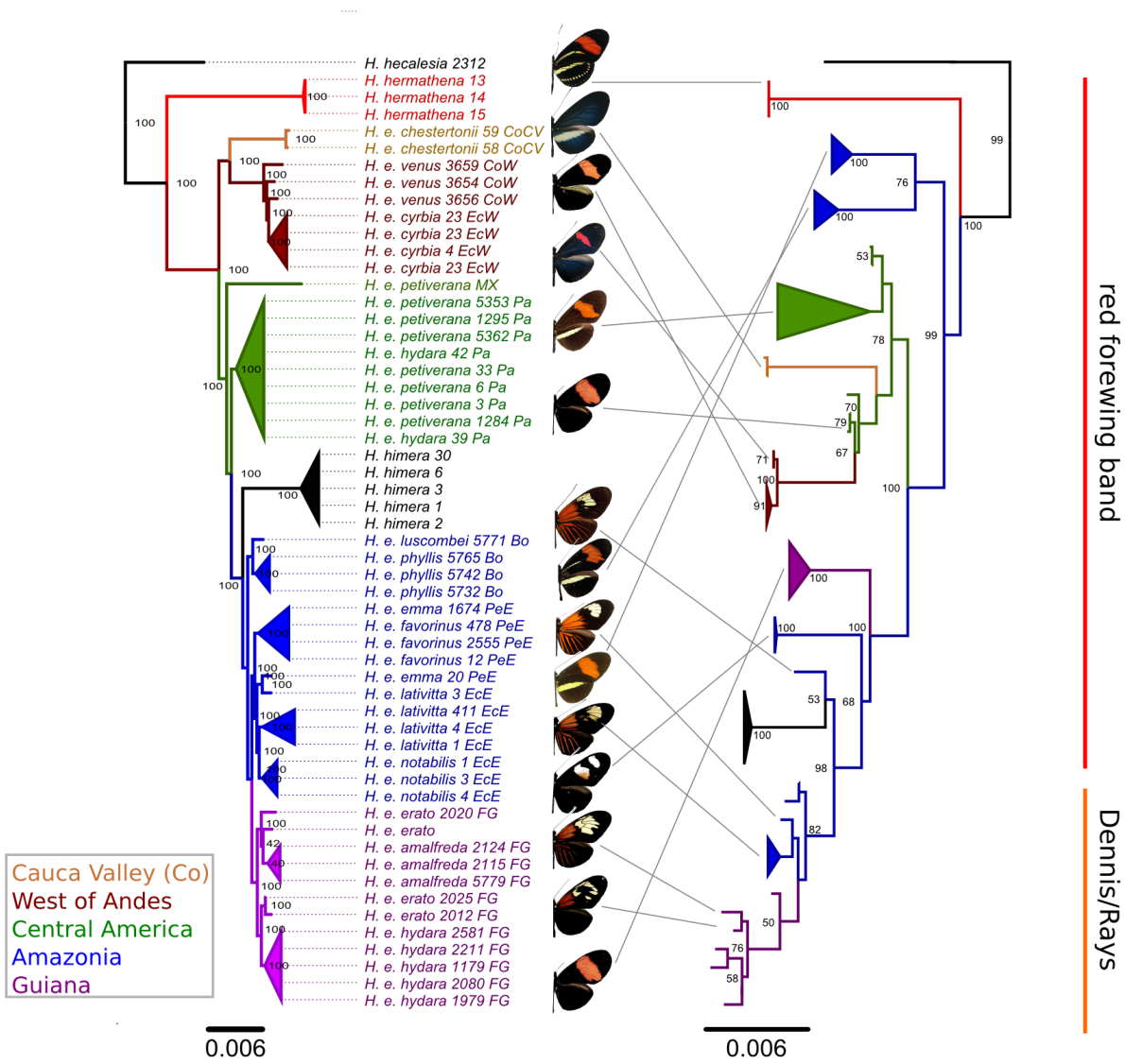


Figure 4.4. Phylogeny at the critical section of the red pattern locus *D* is highly distinct from the history of divergence at the putatively neutral loci. Left: ML tree of the *H. erato* races based on 17 mbp of genome-wide sequence. Notice clustering by geography and full resolution and support. Right: *D* locus alleles cluster primarily by phenotype. Similar results were found previously by Supple and colleagues (2013).

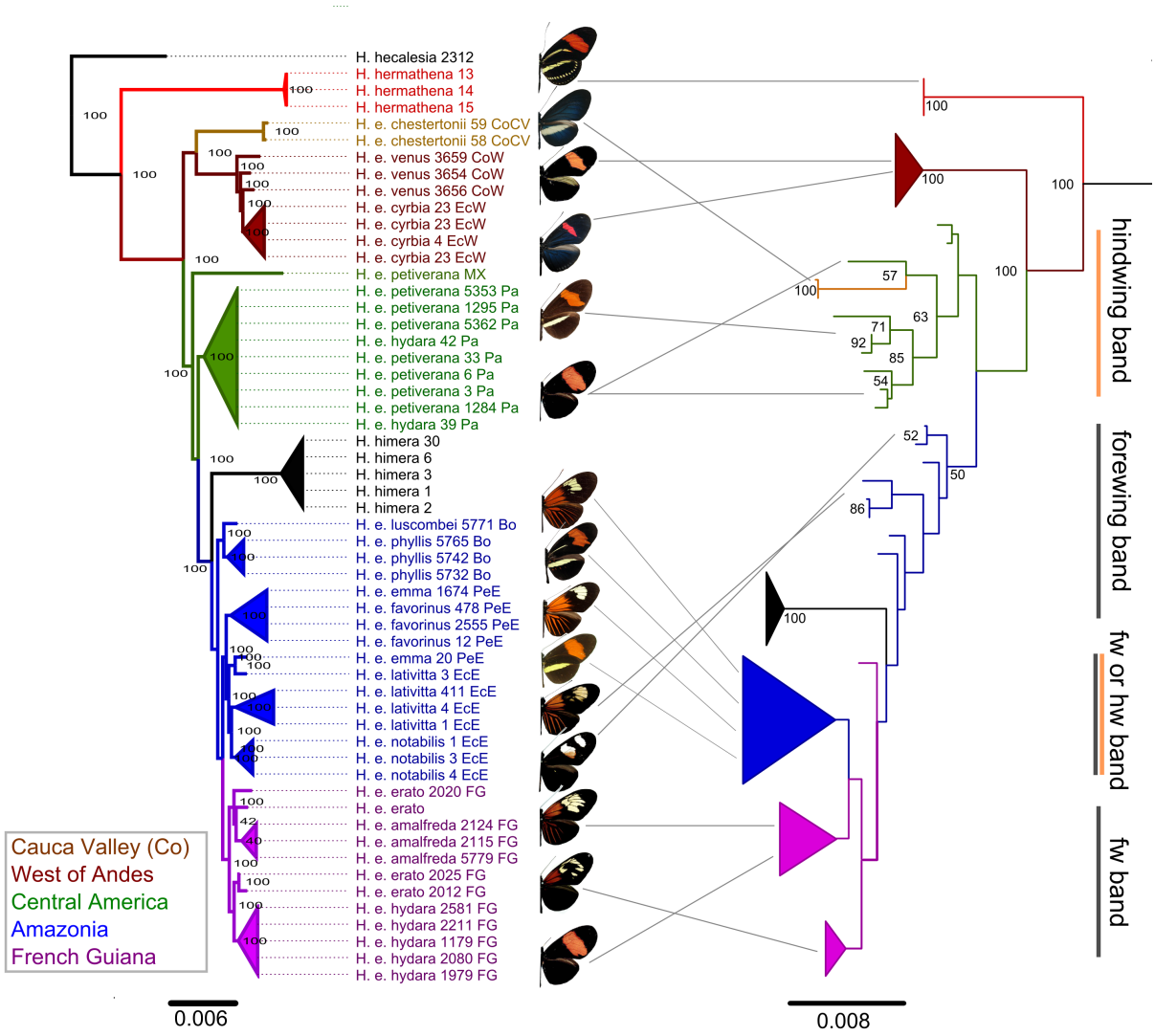


Figure 4.5. *Heliconius erato* alleles of the yellow pattern locus *Cr* show clustering by geography, not phenotype. Only nodes with bootstrap support over 50/100 labeled.

Robust evidence for orogeny-driven population structure in Heliconius erato

The diversification history of *H. erato* races, the largest phenotypic radiation in Heliconiini, has been subject of a prolonged controversy and lies at the core of our understanding of how novel morphologies arise and spread. To investigate this problem in detail and gain insights into the reasons for the *H. erato/hermathena* split I extended the sampling to all available populations and races (Appendix). In contrast to previous efforts (Brower 1994; Quek et al. 2010; Hines et al. 2011; Cuthill and Charleston 2012; Hill et al. 2013), based on genome-wide variation I reconstructed the phylogeographic history of continent-wide colonisation with high support (Fig. 4.4). Resolution of branches between populations was not merely an artefact of enforcing the tree structure on the data, as a split network also shows multiple distinct clades corresponding to geography (Fig. S4.4). The order of branching in the chronogram supports the idea that the ancestral population of *H. erato* existed in the North-West of South America, West of the Andes, after diverging from *H. hermathena* 4.2 MA (95% HPD: 3.0-5.1 MA) (Fig. 4.6, Fig S4.5). The modern Central American lineage split off from the Colombian Andes populations 2.7 MA (2.0-3.5), and later diverged from *H. himera* and the Amazonia/Guianas group 1.9 MA (1.4-2.4). The mitochondrial data are broadly consistent but demonstrate that *H. hermathena* alleles are derived from the Eastern populations of *H. erato*, (Fig. S4.1).

The phylogenetic analysis at the *D* locus confirms and extends the results of Supple et al. (2013), specifically recapitulating the pattern of clustering primarily by phenotype. The red forewing band pattern was ancestral and shared with *H. hermathena*, later spreading across the range, whereas the *Dennis/Ray* morphology characteristic of the Amazonian races (the gap in Fig. 4.1) appeared more recently (Fig. 4.4-5). *Heliconius erato chestertonii*, now considered a separate species by some taxonomists (G. Lamas, unpublished), diverged early in the Colombian Cauca Valley, but it does not have any red patches (Fig. 4.4). The topology

at the *Cr* locus, albeit poorly supported, is congruent with phylogeography of the putatively neutral loci and implies no exchange between similarly patterned races (Fig. 4.5). AMOVA infers that 32.8% of the variation is explained by geography, whereas the yellow wing pattern accounts for only 9.1%. This indicates that there may potentially be other loci playing a large role in the development of the yellow patterns.

DISCUSSION

The concept of homoploid hybrid speciation is of central importance in the intense debates on the mechanism by which new species arise, as it challenges the established notions on the homogenising role of gene flow and the nature of genomic factors leading to speciation in animals (Abbott et al. 2013; Schumer et al. 2014). A specific form of HHS, Hybrid Trait Speciation (HTS), offers one of the key routes for the emergence of new species in sympatry and parapatry through ecological selection on novel combinations of traits (Jiggins et al. 2008). Studies of *Heliconius* butterflies have so far produced some of the best examples of hybrid speciation in the animal kingdom (Mavárez et al. 2006; Salazar et al. 2010; Heliconius Genome Consortium 2012; Pardo-Díaz et al. 2012; Wallbank et al. 2015), spectacularly manifested in composite patterns, although the documented cases are restricted to a clade constituting a fraction of heliconian diversity. In this chapter I test whether *Heliconius hermathena,* belonging to a different branch of the *Heliconius* tree, evolved by either genome-wide admixture or adaptive introgression between *H. erato* ad *H. charithonia* after at least 3 MA of divergence, as is suggested by its composite wing patterns. My results indicate that the species has in fact arisen from a common ancestor with *H. erato* and converged upon the characteristic zebra pattern of *H. charithonia*.

Absence of signal and methodological perils

Protein-coding loci and millions of non-coding sites unequivocally place *H. hermathena* as the sister species to the *H. erato* complex (including *H. chestertonii* and *H. himera*), showing that the discordant signals found by Beltrán and colleagues (2007) are idiosyncrasies of a specific nuclear marker. Although stochastic population-level processes in the ancestral species could obscure the signals in the concatenated analysis of CDS and non-coding loci (Edwards 2009; Thiergart et al. 2014), the coalescent approaches largely account for this problem and identify only two out of 921 genes as plausibly supporting a close relation between *H. hermathena* and *H. charithonia*. This limited signal could result from recent gene flow, as documented between *H. melpomene* and *H. cydno/timareta* (Martin et al. 2013), but the splits between the alleles in both gene trees are relatively deep. The gene *Rotund* is implicated in eye and wing patterning of *D. melanogaster*, neatly fitting the story of a redeployment of eye development pathways on butterfly wings (Reed et al. 2011). *Rotund* regulates *Wingless*, which is involved in wing development and is loosely linked to the *K* locus determining pattern and mate preference of *H. cydno* (Kronforst et al. 2006; Martin et al. 2012).

The ABBA-BABA tests used in this study have proven overly sensitive to noise in the data, making the specificity of statistical tests the key problem in detecting and quantifying admixture. Calculation of *D* and *f* statistics is not a reliable hypothesis-testing device in this case, as statistically significant evidence of gene flow is obtained for nonsensical scenarios, including gene flow from allopatric and phenotypically very different species, which are not known to hybridise in the wild with either *H. hermathena* or its progenitors (Mallet et al. 2007). Critically, I expected that the degree of admixture should vary substantially between parts of the genome, with the highest levels at the pattern loci under strong selection (Pardo-

Díaz et al. 2012), and little gene flow at the sex-linked Z chromosome due to Haldane's Rule, as previously observed in *Heliconius* (Martin et al. 2013) and mosquitoes (Fontaine et al. 2015). Gene flow should also be higher between *H. hermathena* and the parapatric Guianian or Amazonian races of *H. erato* than the allopatric trans-Andean forms. Counterintuitively, the admixture estimates are similarly low but very statistically significant across comparisons, leading to the conclusion that the test produces false positive results, perhaps caused by violating the assumptions on ancestral population structure and effective population size (Green et al. 2010; Martin et al. 2015). The test was primarily designed for closely related ingroup taxa with sympatric and allopatric populations, such as African and European modern humans, with a shared recent history. Many of the assumptions may there fore not be met when applying it to more distantly related taxa. Both putative progenitors are highly structured species (Davies and Bermingham 2002; Hines et al. 2011), and all three species differ dramatically in their population size, from widespread and diverse *H. erato*, through dense but more geographically restricted *H. charithonia*, to very localised and rare *H. hermathena* (Brown and Benson 1977; Rosser et al. 2012). Although the ABBA-BABA calculations are robust to N_e differences of an order of magnitude, specificity of the measures at the extremes is unknown (J. Davey, *pers. comm.*). Alternatively, some of the observed signal may also reflect a degree of gene flow between the ancestral species in the radiation, including the *H. clysonymus* and *H. sara* lineages, although this scenario could only be tested against complex simulations.

Evidence suggests that hybridisation between *H. charithonia* and *H. erato* is a very rare occurrence (Mallet et al. 2007), opening up the possibility of admixture followed by extensive backcrossing to *H. erato*, in which case only a very small proportion of the genome would resemble *H. charithonia* haplotypes. Adaptive introgression driven by selection does in fact appear to be much more common than genome-wide mosaicism (Schumer et al. 2014),

and has been documented in taxa ranging from *Drosophila* (Brand et al. 2013) to mice (Song et al. 2011). Even in taxa where gene flow is rampant, the size of the admixed block responsible for the key adaptation may be minuscule relative to the genome, as recently documented for a beak shape-regulating locus in Darwin's finches *Geospiza* (Lamichhaney et al. 2015). Alas, alack! The story of adaptive introgression of wing pattern loci into *H. hermathena* does not withstand the scrutiny, as none of the tested colour pattern haplotypes from the suspected hybrid cluster with *H. charithonia* (Table S4.2). An intriguing exception is the association of coding sequences from *poikilomeusa*, the key regulator of yellow phenotypes, but the clustering is not statistically significant at the p=0.01 level or upheld by intronic sequences. The marginal similarity of the CDS may be partially due to an adaptive convergence between the two zebra-patterned species, although no tracts of sequence similarity are apparent and the residues identified as being under purifying selection differ in both species. As above, the site-based test of introgression is uninformative about the history of the colour pattern loci due to low number of ABBA-BABA sites and general instability when applied in short windows (Martin et al. 2015).

Reliance on several analyses is essential in verifying plausible instances of admixture, as spectacularly demonstrated in the case of *Xiphophorus* fish. Jones and colleagues (2013) used RAD-seq data to generate a concatenation phylogeny for the 27 swordtail and platyfishes, and based on similarities in long haplotypes proposed two hybrid taxa, including *X. clemenciae*. Independently, Cui and colleagues investigated the exact same taxa with RNA-seq (2013), identified nodes where several loci are incongruent under a Bayesian concordance model (see Chapter 2) and scrutinised the specific branches with ABBA-BABA metrics. Curiously, they discovered as many as seven instances of admixture, but not the two found by Jones et al. (2013), and in a companion paper argued that *X. clemenciae* is not a hybrid (Schumer et al. 2013). Diverse phylogenetic and population genetic analyses (concatenation

and coalescent phylogenies, Riata's test, ABBA-BABA statistics, reticulation networks, gene trees in sliding windows: Fig. 4.3-4, S4.1-4) demonstrate that *H. hermathena* is the sister species of *H. erato* and resembles it across the genome, with little credible evidence for even very localised introgression from the highly diverged *H. charithonia*. Not even recent gene flow from the sister species is inferred (Table 4.2), as expected given the isolation of *H. hermathena* unique grassland habitat. Consistent with a previous general critique of the adaptive introgression model of *Heliconius* pattern evolution (Brower 2013), the superficial similarity to *H. charithonia* is likely an example of convergence by redeployment of an ancestral toolkit, as the diverse *Heliconius* species have evolved modifications to the same set of essential patterning genes (Huber et al. 2015). Sheppard and co-authors (1985) postulate that the ancestral morphology is most likely to survive at the fringes of the distribution. At least two out of the four zebra-patterned species in the genus (*H. charithonia*, *H. nattereri*) are found at the edges of the tropical biomes, and it could be argued that the very unusual riparian habitat requirements of *H. hermathena* in the Amazon and *H. luciana* in the Guianas effectively make them fringe species (Brown and Benson 1977; Brown 1981). The similarity in *poik* CDS provides some evidence for retention of the ancestral variation, although the customary cry for more data is justified here: the two samples of *H. charithonia* and three of *H. hermathena* are insufficient for proper selection and Linkage Disequilibrium analyses, which would throw more light on the history of the *Cr* locus. Based on the evidence presented here I can suggest that *Cr* (yellow) has evolved similarly to the rest of the genome and without extensive introgression found for the *D* (red) locus.

Allopatry, parapatry and divergence in the Heliconius erato radiation

There is a long history of research into *H. erato*, as this massive phenotypic radiation and its nearly-perfect counterpart in *H. melpomene* form the leading example of mimicry-driven intraspecific divergence and interspecific convergence. Brower (1996) used *Cytochrome Oxidase I/II* data to demonstrate a deep split between populations found East and West of Andes, and the first estimate of the arthropod mitochondrial molecular clock rate to infer simultaneous divergence in Pleistocene forest refugia around 200 Kya. Flanagan et al. (2004) found demographic evidence of a population expansion of *H. erato* ~2.5 MA, greatly predating the Pleistocene, and date the divergence of *H. himera* at ~2 MA. Quek and co-authors (2010) sampled an impressive variety of phenotypes across the range of *H. erato* and suggest it originated on the Eastern slopes of the Andes, but the phylogeographic trees reconstructed from AFLPs have negligible bootstrap support. Hill and colleagues (2013) proposed an origin in Central America, although their argument is based purely on mitochondrial evidence.My explicit re-analysis of their AFLPs and Z-linked (*Tpi*) sequences shows no phylogenetic structure (not shown), indicating that the mitochondrial tree is the only source of support for the conclusion drawn by Hill and colleagues. Other work demonstrates that the loci defining the red patterns evolve entirely independently from the rest of the genome (Hines et al. 2011) and the Amazonian rayed patterns are the most recently derived (Supple et al. 2013).

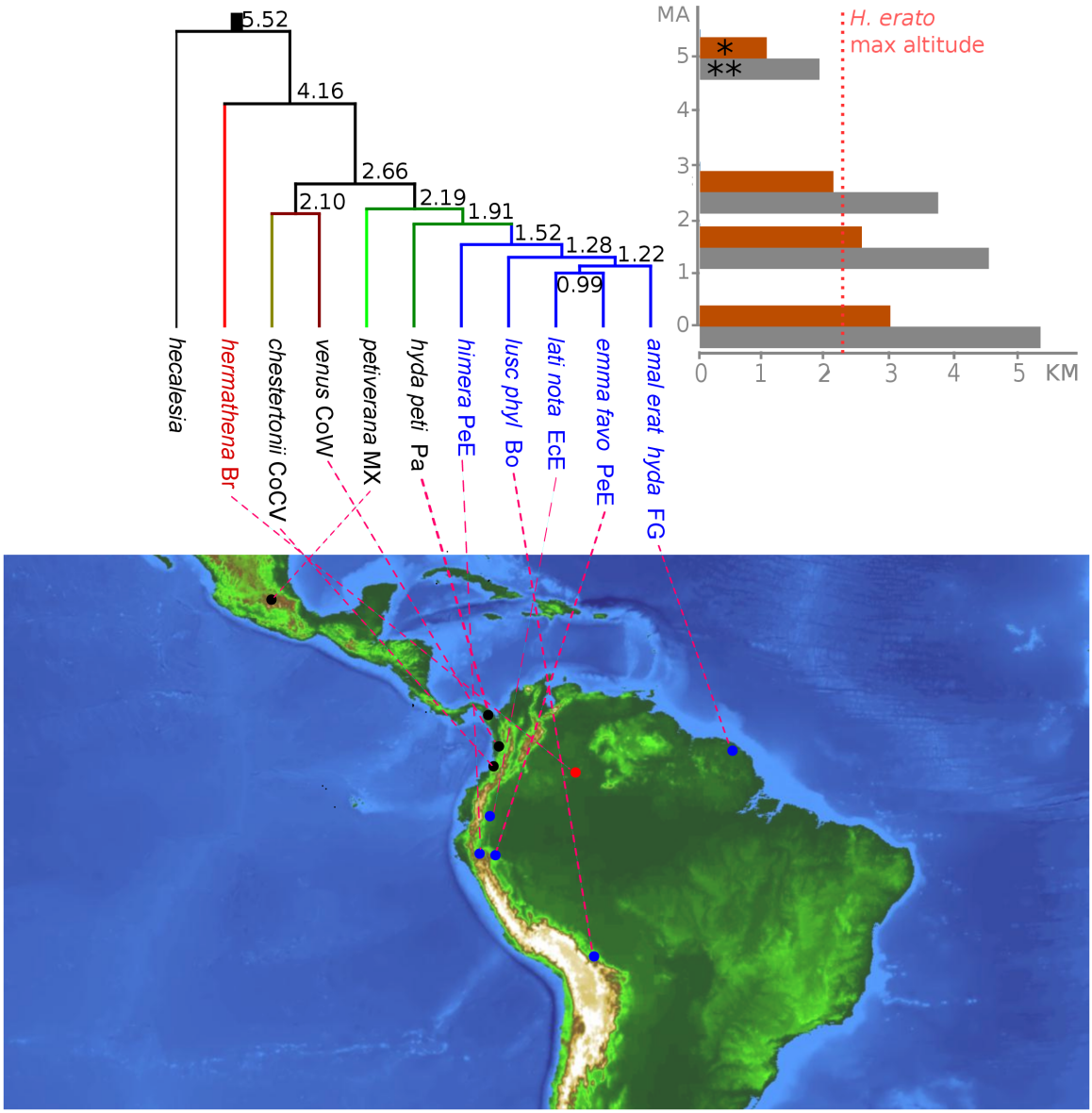


Figure 4.6. The phylogeographic divergence pattern of *Heliconius erato*. Top left: mean divergence estimates from the BEAST analysis reported. Top right: average altitude of Colombian Andes: brown (*) mean altitude of Cordillera Central; grey (**) the highest peak. Mapping in R package *phytools* (Revell 2012) over a NASA topographic map. Andes labeled in pale green-brown-yellow scale corresponding to increasing altitude.

I capitalise on the unprecedented availability of genome-wide data to resolve the history at the neutral loci and propose a novel scenario where the ages of splits between populations are correlated with geophysical changes (Fig. 4.6). The *H. erato/hermathena* split coincided with the reduction in connectivity between the West coast of South America and the rest of the continent during the fast uplift of the Colombian Cordilleras 5-2 MA (Gregory-Wodzicki 2000; Ochoa et al. 2012), which would substantially hinder gene flow. The divergence of *H. hermathena* may have been further reinforced by the reduction of river bank grasslands in the same period, between the disappearance of the large internal Lake Pebas in late Miocene and the full entrenchment of the Amazon and its tributaries (Hoorn et al. 2010). Rock uplift formed the deep Cauca and Magdalena Valleys separating the three Colombian ridges, which reached over 2000 m in average height shortly before 2 MA (Gregory-Wodzicki 2000; Egbue and Kellogg 2012), approaching the upper altitudinal limit of 2300 m for *H. erato* (Brown 1981). At the same time the *H. chestertonii* lineage split off, perhaps colonising the newly formed high-elevation habitat in the Cauca Valley of Colombia, where today it forms a very narrow and localised, bimodal hybrid zone with *H. e. venus*, showing strong assortative mating and low levels of admixture (Arias et al. 2008). The unusually patterned *H. chestertonii* has been recently elevated to the rank of species (G. Lamas, unpublished), and this distinction is supported by genome-wide divergence from *H. e. venus* seen in my data (Fig. 4.4). As pointed out in a study of tree frogs, the dichotomy between sympatric and allopatric speciation on the Colombian slopes is somewhat artificial, since the rising geographical barriers to dispersal have also been novel ecological theatres of speciation (Muñoz-Ortiz et al. 2015).

Consistent with the role of Andes as a barrier, early lineages of *H. erato* would have dispersed simultaneously South and North along the mountains, with the latter branch diverging further into Central American and Eastern populations (Fig. 4.6). The development

of the Isthmus of Panama was perhaps less relevant than previously thought (Hill et al. 2013) if we believe the new studies dating it to mid-Miocene (11-9 MA; Farris et al. 2011). Even in the absence of a continuous land bridge many species of *Heliconius* appear capable of short-distance dispersal between Caribbean islands (Davies and Bermingham 2002). The only evidence for a completely Isthmenian origin comes from the mitochondrial data, which are at at odds with the nuclear evidence (Fig. S4.1). Such cytonuclear discordance is well-known in recently evolved lineages like *H. cydno* (Salazar et al. 2008) and *H. melpomene* (Quek et al. 2010). *Heliconius erato* may have been able eventually to reach the newly formed Amazon Basin through the small depression at the Northern tip of the continent and spread around 1.5 MA. Many peripheral populations became entrenched in the valleys on the Eastern slope, and perhaps in the geoclimatically unique ecosystems of the Shield of Guiana, although the observed patterns may be due to isolation-by-distance and should be verified with samples from central Amazonia. Some hybridisation with *H. hermathena* had likely taken place, as the mitochondrial haplotype of this species clusters with those of Eastern *H. erato* populations (Fig. S4.1).

The importance of general biotope differences in driving *Heliconius hermathena* speciation is especially plausible when compared to *H. himera,* studied extensively as an example of the role of wing patterns and high-elevation habitat in *Heliconius* diversification (e.g. Jiggins et al. 1997; McMillan et al. 1997). Although derived from *H. erato* and occasionally hybridising, *H. himera* shows strong pre-mating isolation (McMillan et al. 1997). My data demonstrate high genetic differentiation of the species, yet this and previous studies (allozymes: Jiggins et al. 1997; AFLPs: Quek et al. 2010) cluster this species with the Eastern races of *H. erato*. The present-day range is predominantly Western but stretches east of the Andes in the Marañon Valley, and collocates with remnants of the Western Andean Portal (the Marañon Gap), a sharp depression in the cordilleras currently reaching only over 2000 m

above the sea level (Hungerbühler et al. 2002), which may function as a corridor between the slopes.

Although my wide geographic sampling and unprecedented number of sites result in a robust phylogeographic pattern, the history of the mimetic pattern loci requires a separate treatment (Hines et al. 2011). As in the previous study based on some of the same samples (Supple et al. 2013), my work establishes that the alleles encoding the rayed patterns in *H. erato* have certainly evolved last. The red band of *H. erato* and *H. hermathena* is therefore most likely encoded by a shared ancestral variant (Fig. 4.4). The key regions at the *Cr* locus are less easily identifiable, and a genetic analysis at hybrid zones shows that the SNPs associated with various wing patterns are often found at some physical distance within the region (Nadeau et al. 2015). Hence it is necessary to investigate the history o the region in greater detail than presented here.

Conclusion

Diverse approaches to the analysis of genome-wide variation in the *Heliconius erato/H. sara* subclade revealed no evidence for the hybrid origin of *H. hermathena.* Combined with the analysis of the *H. erato* species complex, my results suggest that *H. hermathena* speciated through geographic isolation or parapatric adaptation to its grassland habitat. I present a robust phylogeographic hypothesis for *H. erato*, which is required to understand how evolution by dispersal and genetic drift served as a backdrop to the diversification of the patterning loci driven by natural and sexual selection.

PHYLOGENOMICS DEMONSTRATES NO ADMIXTURE IN MOST *HELICONIUS*

A noticeable conceptual shift in modern evolutionary biology has occurred towards recognising the importance and prevalence of porous species barriers between animal species, leading to frequent hybridisation at different taxonomic levels and demonstrating itself as detectable gene flow (Mallet 2005; Abbott et al. 2013; Schumer et al. 2014). Whole-genome sequencing increasingly yields support for this notion and helps to detect and quantify introgression precisely (Twyford and Ennos 2012; Nadeau et al. 2013; Fontaine et al. 2014; Sankararaman et al. 2015; Lamichhaney et al. 2015). Different loci across a genome can very widely in their history, owing to intraspecific gene flow, as well as a range of other processes, including retention of ancient polymorphism and population structure, incomplete lineage sorting, selection, gene shuffling or duplication and loss (reviewed in Chapter 1). This complexity leads to the prediction that phylogenies estimated from various parts of the genome may be largely incongruent (Takahata 1989; Degnan and Rosenberg 2006; Ané 2011). Phylogenomics is therefore much more than an extension of phylogenetics using more data. Rather, it should focus on identifying the incongruences and their sources, thus uncovering the complexities of speciation and adaptation that produce genomic complexity and interfere with simple inference of phylogenies.

The few genomic studies of entire radiations published to date confirm the prediction of high discordance. Analyses at the level of 16 *Anopheles* species (Fontaine et al. 2015) and

all 48 families of birds (Jarvis et al. 2014) found that no individual gene trees, whether based on protein-coding sequences or large sliding windows, are identical with the species tree topology. This phenomenon appears to be partially attributable to incomplete lineage sorting (ILS) in rapidly radiating groups. For example similar patterns are seen in Darwin's finches speciating after the colonisation of the Galapagos archipelago (Lamichhaney et al. 2015), mammals in response to optimal climatic conditions (Hallström and Janke 2010; Song et al. 2012), and birds immediately after the K/T extinction event (Jarvis et al. 2014). Hybridisation between species that are not fully reproductively isolated is also a major factor, extensively documented in population genomic studies of recently diverged taxa (e.g. anatomically modern humans, Denisovans and Neanderthals, Sankararaman et al. 2014; subspecies of the house mouse, White et al. 2009; *H. melpomene* and *H. cydno*, Martin et al. 2013). At least two species of *Geospiza* finches show gene flow in parapatry (Lamichhaney et al. 2014), and the detailed study of *Anopheles* mosquitoes identified gene flow so extensive that 97% of the autosomal genome has a history different from the most plausible order of divergence (Fontaine et al. 2015). Introgression is predicted to be limited at the sex chromosomes owing to their role in reproductive isolation (Lima 2014) and evidence exists for less gene flow at the X chromosomes in the *Drosophila simulans* (Garrigan et al. 2012) and the Z chromosome in the *Heliconius melpomene* group (Martin et al. 2013). In all *Drosophila* (Pease and Hahn 2013) and *Anopheles* (Fontaine et al. 2015) exceptional levels of discordance were found between the autosomal and sex-linked loci, although the interpretation of these patterns is less certain. Nonetheless hybridisation has been considered less frequently than ILS in the studies of radiations so far and poses greater challenges to computational methods (Nakhleh 2013), as demonstrated by the almost opposite conclusions from two methodologically different studies of the same assemblage of *Xiphophorus* swordtail fish (Cui et al. 2013; Kang et al. 2013).

Apart from stochastic gene flow across the species barrier, genome-wide studies have repeatedly revealed unexpected occurrences of adaptive introgression, where natural selection acts to fix the alleles introgressed from a different lineage. A major locus contributing to the variation in the beak shape among the iconic Darwin's finches has been mapped to a haplotype introgressed between multiple species (Lamichhaney et al. 2015). There is also evidence that selection has driven a small block of transcription factor binding sites from the ecological generalist *Drosophila simulans* to the specialist *D. sechelia* (Brand et al. 2013). In perhaps the most spectacular case described to date, it is claimed that selection for insecticide resistance in African mosquitoes intensified by recent malaria prevention efforts, has driven a resistance allele across *Anopheles gambiae* populations and into *A. colluzzi* (Clarkson et al. 2014; Norris et al. 2015).

Lepidopteran genomics

Our ability to study hybridisation still depends on the availability of the appropriate data. Lepidoptera are the second most speciose insect order and provide ample opportunities for phylogenomic analysis at various taxonomic levels. Rapid progress has been made in some subgroups: agricultural utility has justified the work on the silk worm *Bombyx mori* (Bombycidae) (Xia et al. 2004) and the diamondback moth *Plutella xylostella* (Yponeumetidae) (You et al. 2013). Genomics research on butterflies has so far focused on a range of Nymphalidae species with interesting ecological traits, encompassing the migratory monarch *Danaus plexippus* (Zhan et al. 2011), hybridising swallowtails *Papilio glaucus* and *P. xuthus* (Cong et al. 2015; Nishikawa et al. 2015), the icon of metapopulation biology - glanville fritillary *Melitea cinxia* (Ahola et al. 2014), and the tropical Müllerian mimic *Heliconius melpomene* (Heliconius Genome Consortium 2012). Genome-wide resequencing data are slowly becoming available for other families and tribes, recently revealing the

emergence of unique aspects of the lepidopteran developmental processes through an early expansion of homeobox genes (Ferguson et al. 2014) and miRNAs (Quah et al. 2015).

Phylogenomic approaches have been widely adopted and improved by Lepidopteran systematists. For example, RNA sequencing has been used to address the contentious problem of the placement of butterflies among moth superfamilies (Bazinet et al. 2013). Sampling error has been reduced by adding more sites, rather than taxa (Breinholt and Kawahara 2013; Kawahara and Breinholt 2014). Similarly, whole mitochondrial genomes have been demonstrated to generate a fairly reliable approximation to the family-level phylogenies of butterflies (Wu et al. 2014), and provide a first step in resolving the contentious backbone of the lepidopteran radiation with cost-effective taxon sampling (Timmermans et al. 2014).

Although significant progress has been made with the new resources, studies of deep divergences in the Lepidoptera, as well as other insects (e.g. Misof et al. 2014) have so far largely ignored the sources of phylogenetic incongruence. Both ILS and hybridisation have been dismissed with the assumption that large amounts of data will inevitably resolve any uncertainties and conflicts (Kawahara and Breinholt 2014). Ironically, more detailed studies at the level of closely related species have provided some of the best examples of hybridisation in speciation, including the hybrid *Papilio appalachiensis* (Kunte et al. 2011; Zhang et al. 2013) and multiple *Heliconius* species (Salazar et al 2010; Heliconius Genome Consortium 2012; Martin et al. 2013; Nadeau et al. 2013). This appreciation of the importance of hybridisation needs to be better incorporated into phylogenetic studies.

Phylogenomics of Heliconius hybridisation

Due to its diversity at many timescales, well-understood genetics, fast-growing databases of genomic data, and tremendous background of biological knowledge, *Heliconius* stands out as an excellent system in which to study speciation and hybridisation (reviewed by Jiggins et al. 2008; Supple et al. 2014). The natural propensity of *Heliconius* and relatives, especially *Eueides*, to produce hybrids in the wild (Mallet et al. 2007) has generated an early interest in the genetic porosity of the species barrier (Beltrán et al. 2002; Bull et al. 2006), which has since been explored in depth in the *H. melpomene/cydno* subgenus. Although the group has been previously demonstrated to share variation between ecologically divergent species (Kronforst et al. 2006; Nadeau et al. 2012, 2013; Mérot et al. 2013), the work of Martin and colleagues (2013) has created considerable interest with the finding that although *H. melpomene*, *H. timareta* and *H. cydno* are good species, up to 40% of their genomes can be freely exchanged in sympatry, and this process accelerated after an initial period of strong reproductive isolation (Martin et al. 2015b). Specifically, colour pattern loci are especially likely to be shared between species, thus providing a form of standing genetic variation in a strongly selected trait (Heliconius Genome Consortium 2012; Pardo-Díaz et al. 2012). *Heliconius heurippa*, also belonging to the *H. melpomene* clade, remains the best documented (Schumer et al. 2014) case of Homoploid Hybrid Speciation (Salazar et al. 2008), where the adaptive red wing pattern elements introgressed from *H. melpomene* into the *H. cydno* background (Salazar et al. 2010) and which can be recapitulated in the lab (Mavárez et al. 2006). More surprisingly, the Dennis/Ray pattern found in *H. melpomene* and relatives, as well as in *H. elevatus*, is a product of bilateral adaptive introgression (Heliconius Genome Consortium 2012; Wallbank et al. 2015), despite 5 million years of divergence.

Despite these findings, which have challenged widely held views of the nature of species and speciation (Mallet et al. 1998; Jiggins 2008a), it is virtually unknown whether the

phenomena documented in one, relatively recently emerged (1.8-2.6 MA) subclade of *Heliconius*, are typical of the entire genus. Museum collection data suggest that hybridisation happens most frequently among the species in the *H. melpomene/cydno* and silvaniform groups, collectively called the MCS clade, which also can 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). By using the large amounts of recently available whole-genome resequencing data, I investigate the prevalence of hybridisation in the genus, quantify its extent and compare the processes at various loci in the genome. In this chapter, I extend the thinking and methodology of Chapter 4 by applying more advanced techniques for genotyping and phylogenetic inference, while also revisiting some of the technical issues of phylogenetic inference presented in Chapter 2. Critically, I focus on the practical solutions for distinguishing the effects of hybridisation from those of other evolutionary processes, and apply coalescent and network approaches. In the interest of rapid assessment, I pursue the possibility of analysing the data from the entire genus, rather than from quartets of taxa selected *a priori*, as has been the main approach thus far (in *Heliconius*: Heliconius Genome Consortium 2012; Nadeau et al. 2013; Martin et al. 2013; Chapter 4; in Darwin's finches: Lamichhaney et al. 2015; in plants: Eaton & Ree 2013).

I predict that interspecific introgression is a property universal to the whole genus, as could be expected from the phenotypic data (Mallet et al. 2007) and the prevalence of this process in nearly all cases investigated so far. Considering the importance of the wing pattern loci and the fact that they are subject to strong natural selection (Mallet and Barton 1989), I expect that the levels of gene flow will be especially high in those regions, and that they are most likely to show previously undetected cases of introgression. Conversely, the sex chromosomes are expected to show much stronger isolation due to Haldane's Rule. Furthermore, the lower effective population size of the sex-linked alleles on average leads to

faster coalescence, making the gene trees for these loci more likely to show the true history of speciation (Beltrán et al. 2002; Kunte et al. 2011; Fontaine et al. 2015). Finally, I predict that the radical expansion of the number of sites in the dataset as compared to Chapter 2 will lead to dramatic improvements in the quality of tree inference and facilitate making a proper distinction between noise and historical processes (Edwards et al. 2009).

METHODS

Samples and DNA sequencing

This study is based mainly on a selection of short read Illumina data produced for other research (Heliconius Genome Consortium 2012; Kronforst et al. 2013; Martin et al. 2013; Supple et al. 2013; Briscoe et al. 2013; Wallbank et al. 2014). I chose 149 individuals (Appendix), including at least four individuals per species, if available. I maximised the representation of genetic diversity in each species by selecting samples from geographically distant populations, which should enhance coalescent modelling (Edwards 2009). In case of well-studied species like *H. melpomene*, *H. cydno* and *H. erato*, for which dozens of whole genome samples are publicly available, I included one individual from each distinct population and available wing pattern race. I chose the individuals based on the quality of the raw data and the expected depth of sequencing coverage. In addition, I included four individuals from each of the extensively introgressing populations of *H. melpomene* and *H. cydno* studied by Martin et al. (2013), to serve as a positive control for methods detecting hybridisation.

Eleven additional species were sequenced to increase the taxonomic diversity: *H. atthis*, *H. antiochus, H. egeria, H. leucadia, H. peruvianus, H. ricini* (second sample),

H. erato etylus, Eueides aliphera, E. lampeto, E. lineata, E. isabella, E. vibilia, Agraulis vanillae. The new samples expanded the representation of the *H. sara* clade, provided a high-quality outgroup (*Agraulis*) and made possible an evaluation of the hybridisation levels in the sister genus *Eueides* (Mallet et al. 2007). *Heliconius antiochus* and *H. ricini* were collected by the author in the Bakhuis Mountains, Suriname during an expedition in February 2014. These and other samples from the Butterfly Genetics Group collection were preserved in 96% EtOH at -20°.

All samples were sequenced using the Illumina Genome Analyser II, HiSeq 2000 or HiSeq 2500 technology (Appendix: Specimens) with 100 bp paired-end reads, insert sizes ranging between 250 and 500 bp and the sequencing (expected) coverage from 12x to 110x. In case of the 13 new samples, 30-50 μg of thorax tissue was homogenised in buffer ATL using the TissueLyser (Qiagen). DNA was extracted with the Qiagen DNeasy Blood and Tissue kit, purified by digesting with Rnase A (Qiagen) and quantified on a Qubit v.1 spectrophotometer (LifeTechnologies). At the Beijing Genomics Institute (BGI), whole-genome libraries were constructed and sequenced on a HiSeq 2500 with 500 bp insert size, allocating two individuals per 30 Gbp lane for an expected coverage ~50x. Detailed description of the samples can be found in Appendix 4.1.

Short read mapping

I chose to reconstruct individual genomic sequences by mapping to the *H. melpomene melpomene* reference (Heliconius Genome Consortium 2012), rather than by *de novo* assembly. The advantage of this approach is the ease of comparison between various individuals mapped to the same reference, without the need to find homologs between all the assemblies. There is no reference for the W (female-specific) chromosome in *Heliconius*. Instead, I extended the reference to include six sequences of the W genes *Sxl*, *Tra2* and the

CpW2 transposon from silk moth *Bombyx mori*, codling moth *Cydia pomonella* and swallowtail butterfly *Papilio polytes* (Traut et al. 2006; Fuková et al. 2007; Suetsugu et al. 2013; Kunte et al. 2014).

Quality of the raw reads was checked using FastQC v. 0.11 (Andrews 2014). As the reads are generally high quality, I decided that they would not be trimmed *a priori*, but instead "soft clipped" at the mapping stage to avoid loss of data. Reads were mapped to the reference in two stages: a preliminary BWA v. 6 alignment (Li & Durbin 2009) as in Chapter 4, followed by a more thorough remapping with Stampy v. 1.0.18 (Lunter and Goodson 2011). I chose Stampy from dozens of available short read aligners due to its relatively good performance in mapping reads from divergent species (Davey 2013). Aligner parameters were based on previous studies (Chapter 4; Martin et al. 2013) and the age of divergence from the reference (Table 5.1) (Chapter2; Kozak et al. 2015). To find the optimal settings for the most divergent genomes, I systematically experimented with *Eueides lybia* and *Acraea encedon* libraries, testing ranges of BWA mismatch numbers *k* (2, 3), seed sizes *l* (15-35) and Stampy substitution rates (0.1-0.2). Using the *flagstat* evaluation in Samtools v. 1.19 (Li et al. 2009) I found that numbers of mapped and properly paired reads does not change by more than 4% between the extremes, and opted for conservative values (Table 5.1).

Taxon	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

Table 5.1. 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.

Individual raw alignments were sorted and indexed with Samtools, and duplicate reads from the same molecule were removed with Picard v. 1.112 (Fennell 2010). I used the *IndelRealigner* in the Genome Analysis Toolkit v. 3.1 (GATK) to fix the mapping inconsistencies around the insertions and deletions detected by the RealignerTargetCreator module (McKenna et al. 2010; DePristo et al. 2011). Realignment was performed separately for each individual to limit the computation time and preserve sensitivity (DePristo 2014).

Genotyping pipeline

Development of tools to genotype individuals accurately based on read mapping is a fast-growing area of bioinformatics, but most of the software is designed for the analysis of human datasets, which include more consistently processed and much less genetically variable samples. Genotyping more genetically variable populations is not well-understood and the tools need to be examined carefully (Greminger et al. 2014). I chose the Genome Analysis Toolkit (GATK) and compared the speed and accuracy with previously used tools and parameters (e.g. Martin et al. 2013, Supple et al. 2013, Nadeau et al. 2014) against the

alternatives within the GATK framework. The previously used UnifiedGenotyper, which genotypes SNPs and indels separately, has been recently superseded by the HaplotypeCaller, which improves accuracy by modelling the two types of variation simultaneously, as the incidence of SNPs increases closer to indels (van der Auwera et al. 2013). I tested the performance of the new tool on a set of 11 *H. melpomene/cydno* samples and found a predicted computation time on the order $13.5*10^5$ CPU-hours, compared to <240 CPU-hours using the UnifiedGenotyper. The exact scheme for parallelising the tasks across CPUs did not matter, although using Advanced Vector Extensions (AVX) on AMD Opteron 6380 processors unexpectedly inflated the run time by ~40%.

I therefore proceeded with the UnifiedGenotyper and optimised its parameters on a sample of six species, increasing the default values intended for human data: Indel Heterozygosity (default: 0.000125; alternative: 0.001, 0.01, 0.1), Heterozygosity (0.05, 0.1, 0.15), and the maximum fraction of reads with deletions (0.05, 0.5, 1.0). The results were contrasted with the Vcftools v. 1.12 *compare* function (Danecek et al. 2011). Changes to any of the priors altered the total number of SNPs by no more than 1% and thus all the samples were genotyped using the defaults. After the calculation of Per-Base Alignment Qualities (BAQ) (Li et al. 2009), SNPs were called at sites with coverage >4x and quality >20. These are very liberal values, but I deemed them appropriate for a phylogenetic analysis, where a few individual SNPs are not viewed as evidence in isolation. Genotypes were calculated jointly for all individuals of the same species, as a compromise between (i) more accurate modelling of variant distribution, which requires a larger number of samples, and (ii) calling sensitivity, which is maximised for single samples. Although recommended, Base Quality Score Recalibration (DePristo et al. 2011) was not performed due to lack of a suitable panel of confirmed variants that could be used as a reference. The final Variants Calls Files (VCFs) for all 49 species were merged using Bcftools v. 1.0 (Li et al. 2009) and the quality of calls for

each of the 147 individuals was assessed with the Bcftools *stats* function and an in-house Python script shared by Simon Martin (evaluateVCF-03.py; https://github.com/simonhmartin). Due to low coverage (~3x) across the genome, I excluded the *H. doris* sample 8684 and the distantly related *Acraea encedon*, using the higher quality *D. phaetusa* and *A. vanillae* as outgroups.

Exome alignments and gene trees

As in Chapter 3, I chose the protein-coding genes as a subset of the genome that can be effectively treated as discrete and largely non-recombining markers for multilocus phylogenetics. Gene duplication and loss can have a negative impact on this approach (Maddison 1997; Edwards 2009), when short reads from one copy of a duplicated gene map to the other paralogue. I minimised this problem by narrowing my dataset to the genes known to be 1:1 orthologs between *H. melpomene*, *Danaus plexippus* and *Bombyx mori*, as determined in an OrthoMCL analysis (Heliconius Genome Consortium 2012; W. Palmer, *pers. comm.*). I assumed that single-copy genes conserved between these distantly related lepidopteran lineages (diverged 90 and 117 Myr, respectively) are unlikely to be duplicated or lost during the relatively recent divergence of Heliconiini.

The 425 Gb VCF was converted to the lighter tabular "calls file" format using scripts written by Simon Martin (parseVCF-0.2.py) (Martin et al. 2013). 6848 autosomal and 416 Z-linked single-copy gene alignments were extracted and converted to the *fasta* format (gene_fasta_from_reseq.py, calls_to_seq.py), excluding genes found on five scaffolds linked to the colour patterns. Alignments were pruned automatically to avoid artefacts. TrimaAl v. 1.2 was used to remove the taxa for which 50% of residues did not overlap with at least 50% of the other sequences (Capella-Gutiérrez et al. 2009), and uninformative fast-evolving regions were excluded based on Block Mapping and Gathering with Entropy (BMGE)

(Criscuolo and Gribaldo 2010). Individual gene trees were estimated under the GTR+Γ in FastTree v. 2.1, an approximate Maximum Likelihood program shown to perform orders of magnitude faster than standard implementations without loss of accuracy (Price et al. 2010; Liu et al. 2011). Nodal support was estimated as parametric aLRT values (Anisimova and Gascuel 2006).

Incongruence in the data

The agreement between the gene trees was assessed in two complementary ways. First, I calculated the average normalised Robinson-Foulds distance (Robinson and Foulds 1981) between all pairs of gene trees in PAUP* v. 4 (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 selected based on the quality of their genotypes to maximise gene tree completeness (labeled in Appendix). Second, to identify the regions of especially large incongruence, 50% Majority Rule consensi were calculated from the trees pruned to 57 taxa. The relative support for branches leading to resolved nodes was evaluated in terms of the novel information criteria (IC/ICA), which compare the support for a branch with support for alternative groupings (Salichos and Rokas 2013; Salichos et al. 2014). I accounted for the effect of poor resolution in trees built from short alignments by repeating the procedure with the 1000 best-resolved gene trees.

NeighborNet split networks (SNs) based on all exonic, biallelic SNPs were reconstructed separately for each major clade (*H. melpomene*, silvaniforms, *H. doris*, *H. wallacei*, *H. erato*, *H. sara*, *Eueides*) in SplitsTree v. 4 (Huson and Bryant 2006), using LogDet distances calculated from the matrix of 122,913 autosomal SNPs described above. SNs do not offer a clear quantification of reticulation and are not informative about its source,

but visualise any departures from the bifurcating tree due to either biological complexity or modelling inadequacies (Steel 2005).

Species trees

In the absence of known W chromosome markers, whole-mitochondrial sequences were analysed separately as a proxy for the history of the matriline. To avoid artefacts of rate heterogeneity, an optimal partition scheme of tRNAs, rRNAs, CDS codon positions and non-coding sequences was determined with PartitionFinder v. 1.1 (Lanfear et al. 2012). The tree was estimated under the GTR+Γ model in RAxML v. 8 (Stamatakis 2014) with 1000 bootstrap replicates. The possibility of recombination was tested with a Phi test on 100 bp sliding windows in SplitsTree.

The first step in my analysis of the chromosomal data was a traditional "total evidence" estimate of the species tree based on a supermatrix of concatenated data. I joined the single-copy Z-linked genes in Geneious v. 7 (Biomatters Ltd) and estimated a tree under the GTR+Γ model in RAxML. As the autosomal CDS matrix is very large (10,003,871 bp per individual), I used only the 122,913 exonic, biallelic, non-singleton, 100% complete SNP selected sites with the GATK *SelectVariants* function. The ML tree and 100 bootstrap replicates were estimated after an ascertainment bias correction in RAxML (Stamatakis 2014).

To find the approximate times of lineage splits, a Minimum Evolution tree was estimated from the autosomal SNPs with a LogDet correction in MEGA v. 6 (Tamura et al. 2013) and the branch lengths were ultrametricised by relative rate comparison with RelTime (Tamura et al. 2012), constraining the age of the *Heliconius-Eueides* and *Heliconius-Agraulis* splits between (17.0, 20.0) MA and (25.0, 28.0) MA, respectively (Wahlberg et al. 2009). The LogDet model accounts for the apparent heterotachy (a lower rate of substitution; Fig. 5.1) in the MCS clade (Steel 2005).

Evidence for a large amount of incongruence between gene trees led me to use a Multispecies Coalescent species tree method as discussed in Chapter 2. MP-EST v. 1.4 accounts for Incomplete Lineage Sorting (ILS) by maximising a pseudo-likelihood function over the distribution of taxon triples, extracted from the gene tree topologies. A phylogeny with branches in coalescent units is produced (Liu et al. 2010). The Z and autosomal trees were inferred on the STRAW server (Shaw et al. 2013) and triplet distances between gene trees and the species tree were calculated as a measure of discordance similar to and serving as a mathematically distinct complement to the RF distance. Bootstrap support was evaluated by repeating the analysis 100 times with a random sample of 500 gene trees. Individual tips were assigned to species *a priori*. Based on previous studies (Quek et al. 2010; Nadeau et al. 2013; Supple et al. 2013), I treated the Western, Eastern and Guianian clades of *H. melpomene* and *H. erato* as distinct lineages.

Modelling hybridisation

Following the identification of nodes affected by incomplete lineage sorting with MP-EST, I used network approaches to tackle the more complex problem of distinguishing ILS from hybridisation. Only a few recently proposed algorithms purport to disentangle these processes in cases where more than four species are considered (Nakhleh 2013). I initially attempted to analyse a subset of 500 randomly chosen autosomal trees with the algorithms that detect reticulations due to hybridisation from the topologies of individual loci. The method of Huson (Huson et al. 2005) requires priors that are not explained clearly, whereas fitting a Maximum Parsimony network with the algorithm of Yu et al. (2013) with only 18 lineages and one reticulation took over two weeks to complete, Therefore I used faster techniques that integrate over the history of individual sites in the genome instead.

Next, I applied the Ancestral Recombination Graph (ARG) approach in TreeMix v. 1.2

(Pickrell and Pritchard 2012). This method infers the relations between *a priori* specified populations from allele frequency data under the assumption of random genetic drift, and subsequently identifies pairs of taxa that share a larger than expected proportion of allelic variation. Allele frequencies were computed in PLINK! v. 1.07 (Purcell et al. 2007; Purcell 2009), with individuals assigned to terminal taxa as for MP-EST. I used the default function for frequency and genotyping pruning to account for LD between SNPs, thus excluding correlated observations. ARGs were fitted (a) with or without the correction for small sample sizes; (b) allowing up to zero, 10 or 20 migration events; (c) gradually reducing the data to use *Agraulis*, *Eueides* or *H. aoede* as the outgroup. All results were plotted with an auxiliary R script included with TreeMix (plotting_functions.R) (Pickrell and Pritchard 2012).

Method	Program	Data	Goal
Max Likelihood	RAxML v.8	SNPs	Infer a bifurcating tree
Min Evolution, LogDet model	MEGA v.6	SNPs	Bifurcating tree: account for heterotachy
RelTime	MEGA v.6	Tree branch lengths	Ultrametricise the bifurcating tree
NeighborNet split network	SplitsTree v.4	SNPs	Visualise poor fit to the bifurcating tree: reticulation + systematic error
Ancestral Recombination Graph	TreeMix v.1.2	SNPs	Reticulation network: find hybrids
Reticulation network	SplitsTree v.4, PhyloNet v.3	Gene trees	Reticulation network: find hybrids
Multispecies Coalescent	MP-EST v.1.3	Gene trees	Bifurcating tree: account for ILS
Majority Rule Consensus	RAxML v.8	Gene trees	Multifurcating tree: find incompatibilities between loci

Table. 5.2. Not all networks are inferred equal. Distinct phylogenetic approaches in this study aimed at uncovering specific aspects of the evolutionary process.

Introgression at the colour pattern loci

An essential aspect of studying hybridisation is to distinguish genome-wide admixture from localised introgression of specific genes under selection (Hines et al. 2011; Pardo-Díaz et al. 2012). I considered the history of the five major loci associated with aposematic wing patterning separately (Table 5.5). Each scaffold alignment was partitioned into windows of 20 kbp, sliding by 10 kbp, and accepting only windows with data for a minimum of 1000 bp (SM script sliPhy3.py). The topology for every window was reconstructed with FastTree, tested for significant differences from the whole-genome coalescent 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 red control loci of *H. erato* found by Supple et al. (2013), I aligned the *B/D* scaffold HE670865 with the *H. erato B/D* BACs (Papa et al. 2008) in mLAGAN (Brudno et al. 2003).

Hybridisation in the Heliconius erato clade

The position of *H. hecalesia* and the related *H. clysonymus* clade is uncertain in the supermatrix and coalescent phylogenies (Fig. 5.1-5.4). The analysis of colour pattern loci reveals an unexpected pattern: clustering of *B/D* sequences from *H. clysonymus*, *H. hortense*, *H. hecalesia* – the three species displaying an unusual, thick hindwing red band (Fig. 5.7). I investigated the possibility of admixture in this group by PCA, focusing on 39 samples from the *H. erato/sara* clade. The R package *adegenet* (Jombart and Ahmed 2011) was used to calculate the eigenvalues, plot the PCs and estimate admixture percentages in individual genomes using Discriminant Analysis (Jombart et al. 2010). In order to date the introgression, I concatenated the specific 90 kbp of aberrantly grouping sequence at the *B/D* scaffold (HE670865: 320-380 and 410-440 kbp), built an ML tree and dated the splits with RelTime as above.

RESULTS

Sequencing and genotyping

New WG Illumina data were produced successfully for 11 species, extending the sampling of the *H. sara* clade and *Eueides*. Library construction failed for the *E. lineata* and *H. ricini* samples due to non-DNA contamination. 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.

Mapping to the *Hmel1* reference produced high quality alignments for 144 individuals in 48 species, including 40/45 species of *Heliconius*, 6/12 *Eueides* and the monotypic *Agraulis vanillae* and *Dryadula phaetusa*. Most subgenera of *Heliconius* (as defined in Brown 1981 and Chapter 2) 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 (listed in Appendix: Specimens).

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 approximately 5% of the *H. melpomene* genome (Heliconius Genome Consortium 2012), the data obtained for these more distantly related species is mostly non-coding sequence. Nothing was recovered for the putative W chromosome markers. 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) that manifested itself in poor quality of inspected alignments.

The quality of the mapping depends not only on relatedness: samples with expected coverage (total bases sequenced divided by reference length) >20x generally produced better results, presumably due to tiling of reads (Appendix). Visual inspection of dozens of randomly selected CDS alignments showed that even the relatively recently diverged silvaniform clade samples did not always produce full, high quality sequences if Illumina sequencing was performed to 15x or lower expected coverage (e.g. Kronforst et al. 2013). The results were excellent for the 11 newly produced samples and others with expected coverage >40x.

The number of SNPs follows trends similar to those described above, with the highest number of credible variants (Phred QC>20, minimum coverage 4x) found for the MCS clade (Table 5.3). 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, $11.5*10^6$ private variants were discovered just in the *H. melpomene* samples, and private SNPs constituted 30% of the total. 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).

Extensive experimentation with settings of the genotyping pipeline demonstrated that the prior values in the predominantly Bayesian tools do not have a strong impact on the outcome (<4% difference in the number of reads mapped, <1% SNPs called), as reported previously in a pedigree study (Ness et al. 2012). The ultimate limiting factor in genotyping was the computational time, which differed dramatically between the tools. Mapping short reads with BWA as described in Chapter 3 requires under 100 CPU-hours per sample, but recovers only <5% of the sequence, even for moderately diverged taxa. Addition of the sensitive alignment in Stampy inflates the time per sample to 400-1000 CPU-hours, depending on the number of reads. At the genotyping stage the time required to use the *HaplotypeCaller* becomes so dramatically large (approximately $13.5*10^5$ CPU-hours for 11 samples) that only the less sophisticated *UnifiedGenotyper* could be used (<250 CPU-hours), even given the high computational power available.

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
Acraea	1	1	1.51	16,227,689 (14.85%)	3,389,626 (3.10%)	18,077,966	16,080,893	18,077,966	1.51
TOTAL**	**48**	**145**	**17.4 (3.43-100.01)**	**n/a**	**n/a**	**126,865,683**	**90,646,525**	**38,070,723**	**1.29**

Table 5.3. 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. **Excluding *Acraea*.

Gene alignments and trees

6848 single copy genes were found on the autosomes (excluding the five colour pattern scaffolds) and 416 on the Z sex chromosome. Although I did not verify it, those genes are likely to include mostly the conserved set of core lepidopteran CDS (Ahola et al. 2014). Alignment trimming with TrimAl and BMGE had very little effect, on average removing fewer than one individual and six sites per alignment (Table 5.4), although it helped to eliminate poorly mapped sequences in some cases. The smallest alignment contained 86 taxa. Most alignments were of a length appropriate for phylogenetic reconstruction (mean autosomal gene length 1387 bp) and of good quality (4% missing data). Most gene trees were well-resolved, with the average support for interspecific nodes equal 41.8 (out of possible 56). 98.1% autosomal and 97.6% Z-linked alignments contained an *A. vanillae* sequence to be used as an outgroup.

Parameter	Autosomal	Z-linked
# alignments	6848	416
# *Agraulis*-rooted alignments	6724	410
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

Table 5.4. Basic statistics for the autosomal and Z-linked protein-coding gene alignments. Relatively short sequences and uninformative sites were removed. Range of lengths after trimming in parentheses.

Mitochondrial phylogeny

The whole-mitochondrial alignment was automatically divided into six optimal partitions, roughly corresponding to functional classes expected to evolve under similar substitution rates: e.g. one partition for 44/46 of the relatively conserved tRNA genes. No evidence of mitochondrial recombination was found in the Phi test (p>0.1). The topology of the tree agrees with that presented in Chapter 2 and is strongly supported, with 40/49 interspecific nodes recovered in 0.95 or more bootstrap replicates (Fig. S5.1). Among the poorly supported nodes are the "primitive" clades of *Neruda*, *H. wallacei* and *H. doris,* as well as *H. ricini*, *H. sara* and *H. antiochus*. Consistently with Chapter 3 and previous analyses (Quek et al. 2010; Hill et al. 2013), the *H. hermathena* and *H. himera* lineages are nested within *H. erato,* which also shows a clear division into four biogeographic groups (Mexico, West of the Andes, Amazonia, Guiana), with the Mexican sample branching off first. In the MCS clade multiple instances of paraphyly are observed. *Heliconius pardalinus ssp. nov.* and *H. elevatus* form a monophyletic clade to the exclusion of *H. p. sergestus*, even though all the *H. pardalinus* samples come from the same locality, whereas *H. elevatus* was collected widely in Ecuador and Peru. Notably, the unusually patterned *H. hecale clearei* from Venezuela appears distinct from the *H. hecale* individuals collected in Peru and Central America. Albeit with negligible support, the sequences of *H. timareta* group with *H. pachinus,* and not *H. heurippa*, which is nested within *H. cydno*. This is counter to the nuclear patterns (Martin et al. 2013; Nadeau et al. 2013). As observed previously (Salazar et al. 2008), the mitochondria of *H. m. melpomene* (sample "hmm6") from the Magdalena Valley in Colombia are introgressed from *H. cydno*. *Heliconius melpomene* sequences segregate into the usual three clades, although the monophyly of the Amazonian group with the rest of the species is not supported. This may reflect an affinity of the Eastern races, sampled along the Andes, with the peripatric *H. timareta*.


Eueides isabella
Eueides lampeto
Eueides vibilia
Eueides aliphera
Eueides tales
Eueides lybia
H. telesiphe
H. hortense
H. clysonymus
62
H. peruvianus
H. charithonia
H. ricini
H. eratosignis
H. demeter
H. leucadia
H. sara
H. antiochus
H. congener
H. eleuchia
H. hewitsoni
H. sapho
H. hecalesia
H. hermathena
H. erato chestertonii
H. erato cyrbia
H. erato venus
H. erato petiverana Pan
H. erato petiverana Pan
H. erato hydara Pan
H. erato petiverana Pan
H. erato petiverana Mex
H. erato phyllis
H. erato luscombei
H. erato hydara FrGui
H. erato erato
H. erato amalfreda
H. erato favorinus
H. erato notabilis
H. erato etylus
H. erato emma
H. erato lativitta
H. himera
H. aoede
H. egeria
H. burneyi
H. wallacei
H. hecuba
H. doris
H. hierax
H. xanthocles
H. besckei
H. numata
H. ismenius
H. ethilla
H. atthis
H. hecale
H. pardalinus sergestus
H. elevatus
H. pardalinus ssp. nov.
H. timareta
H. heurippa
H. cydno alithea
H. cydno zelinde
H. cydno hermogenes
H. cydno cydnides
H. cydno yepmeri
H. cydno chioneus
H. cydno chioneus
H. cydno chioneus
H. cydno cordula
H. cydno chioneus
H. cydno galanthus
H. pachinus
H. melpomene melpomene REFERENCE
H. melpomene melpomene Col
H. melpomene vulcanus
H. melpomene cythera
H. melpomene melpomene Pan
H. melpomene vulcanus
H. melpomene melpomene Pan
H. melpomene rosina
H. melpomene rosina
H. melpomene rosina
H. melpomene rosina
H. melpomene rosina
H. melpomene melpomene FrGui
H. melpomene meriana
H. melpomene thelxiopela
H. melpomene melpomene FrGui
H. melpomene melpomene FrGui
H. melpomene melpomene FrGui
H. melpomene amaryllis
H. melpomene ecuadorensis
H. melpomene malleti
H. melpomene malleti
H. melpomene aglaope
H. melpomene amaryllis
H. melpomene amaryllis
0.02

Figure 5.1. 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*. Together, the last three groups form the *melpomene/cydno*/silvaniform (MCS) clade.

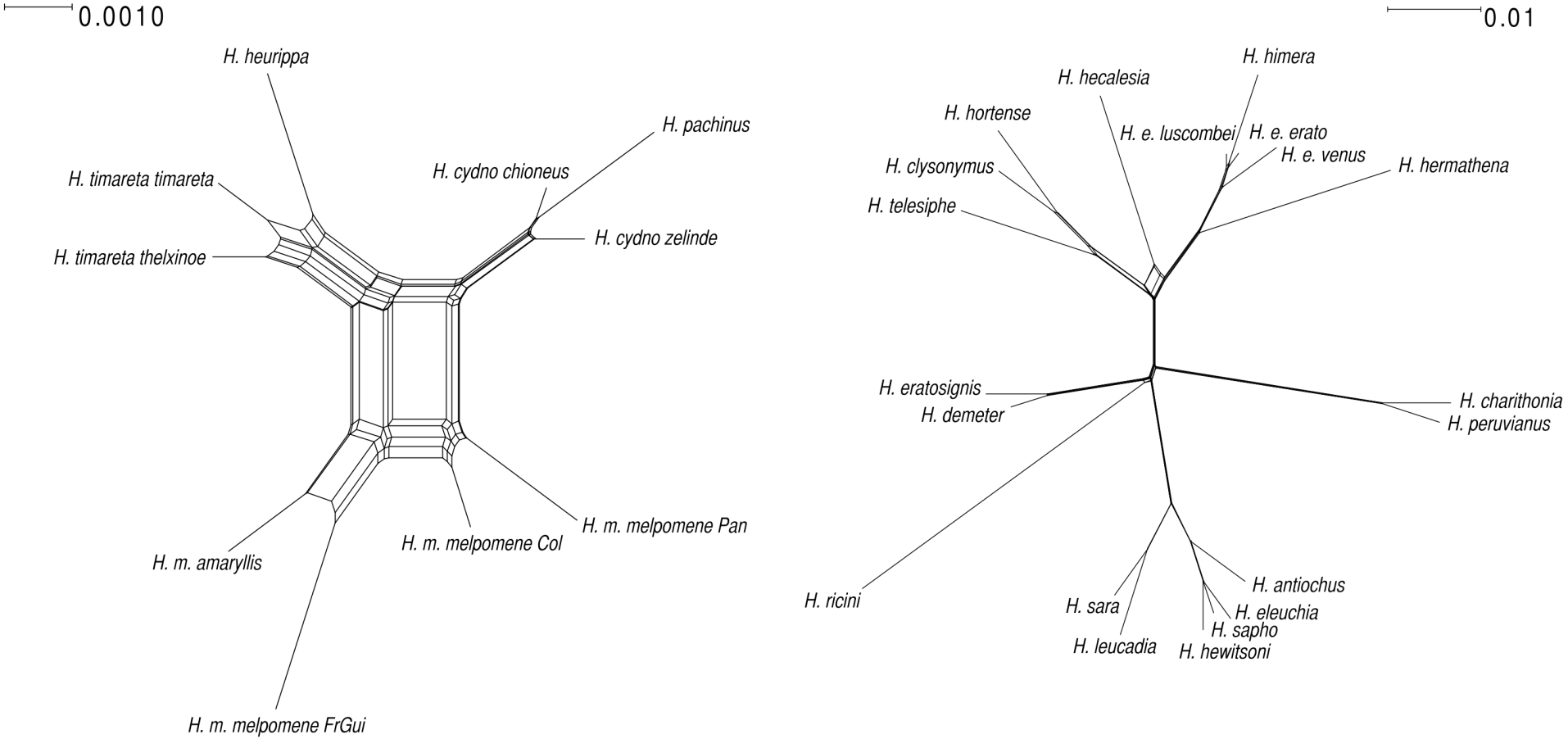
Autosomal supermatrix phylogenies

Filtering for autosomal exome sites with biallelic, non-singleton SNPs without any missing data produced a 122,913 bp matrix. The corresponding Maximum Likelihood tree is nearly completely resolved with very high bootstrap support (Fig. 5.1). The only exception is the uncertain placement of the *H. telesiphe/hortense/clysonymus* clade *vis-á-vis* the *H. erato* and *H. sapho* groups. As in the mitochondrial tree (Fig. S5.1), departures from the supermatrix topology in Chapter 2 (Fig. 2.1) occur at most nodes with poor support in the previous work, including: position of the *H. telesiphe* group; relations in the *H. sapho* clade; relative placement of *H. wallacei* and *H. doris* groups; position of *H. besckei*. SNPs generated in this study, RAD-Seq (Nadeau et al. 2013) and AFLP data (Quek et al. 2010; Arias et al. 2014) agree about relations within *H. melpomene* and *H. cydno*, including the nested position of *H. pachinus*. The races of *H. erato* appear less resolved than in Chapter 4, although contrary to previous studies (Quek et al. 2010) more separation is observed among the races of *H. erato* than *H. melpomene*.

Strikingly, the branches within the MCS clade and its sister *H. doris* group appear two- to four-fold shorter than in the rest of the tree. Smoothing the branch lengths with the

model-free RelTime algorithm generates a chronogram roughly consistent with Chapter 2 (Fig. S5.2). The means are typically similar to those recovered under the relaxed molecular clock model, although the most recent branches (e.g. *H. erato* individuals) appear longer than expected due to the properties of the Minimum Evolution algorithm (K. Kozak, unpublished). The general dating framework for *Heliconius* is thus upheld, although very large errors surround the mean estimates of age. Notably, the mean values show that many of the poorly supported splits occurred within short intervals, e.g. approximately 650,000 years separating the appearance of the *H. sapho*, *H. telesiphe* and *H. erato* clades; 150,000 years between the emergence of the *H. sapho* and *H. ricini* lineages; and only 50,000 years separating *H. hewitsonii/sapho* from *H. congener/eleuchia.*

Figure 5.2. Variable degree of departure from the bifurcating tree model. Split networks based on the autosomal exonic SNPs, where edge lengths correspond to pairwise genetic distances under the LogDet correction (scale bars in number of substitutions per site). Greater width of a rhomboid indicates a greater conflict. Left: the *H. melpomene/cydno* clade. Right: the *H. erato/sapho* clade. Note: networks are not drawn to scale.


0.0010
H. heurippa
H. pachinus
H. timareta timareta
H. cydno chioneus
H. cydno zelinde
H. timareta thelxinoe
H. m. melpomene Pan
H. m. melpomene Col
H. m. amaryllis
H. m. melpomene FrGui
0.01
H. himera
H. hecalesia
H. hortense
H. e. luscombei
H. e. erato
H. e. venus
H. clysonymus
H. hermathena
H. telesiphe
H. eratosignis
H. charithonia
H. demeter
H. peruvianus
H. ricini
H. antiochus
H. sara
H. eleuchia
H. sapho
H. leucadia
H. hewitsoni

Incongruence

NeighbourNet networks constructed from the autosomal data reflect varied amounts of reticulation across the clade (Fig. S5.3). Rhomboidal projections, which indicate departures from a strictly bifurcating tree (Kloepper and Huson 2008), are noticeable between the nodes linking the major subgenera. Closer inspection reveals greater reticulation and shorter branches in the *H. melpomene/cydno* than the *H. erato/sapho* clade (Fig. 5.2), but the reverse is true for the Z chromosome (Fig. S5.6). Intriguingly, known patterns of gene flow are evident in these data (Martin et al. 2013), with the genetic difference between the Western *H. melpomene* and sympatric *H. cydno* only about twice as much as the distance between *H. cydno* and its sister species *H. timareta* and *H. heurippa* (Fig. 5.2). The same relation holds for the sympatric Eastern *H. melpomene* and *H. timareta*. Additionally, the ambiguous positions of the Guianian and Colombian *H. melpomene* are apparent. In contrast, the *H. erato* clade is relatively well-resolved, except for a reticulate vortex at the base of *H. hecalesia* (Fig. 5.2).

The topologies ofML trees for individual genes varied hugely. The mean symmetrical (Robinson-Foulds) pairwise difference in topology was 0.745 for the autosomal and 0.699 for the Z loci. When only one individual per species or distinct lineage (57 tips) was included, the RF was still respectively 0.234 and 0.192, demonstrating that almost a third of the RF difference is explained by the disagreement in interspecific relations. When compared against the species tree, the gene trees showed a departure of 0.15-0.2 with similar differences between the Z and the autosomes (Fig. S5.4).

A striking illustration of the conflict between individual autosomal gene trees is provided by the 50% Majority Rule Consensus (Fig. 5.3), where only 26/56 nodes are resolved, and many among them have a low Information Criterion (IC/ICA) support value, indicating that there is one (IC) or multiple (ICA) alternative nodes found at high frequency

(Salichos and Rokas 2013). The relative tree certainty for this tree (TCA) is 0.322 on a 0-1 scale. If only the 1000 most-resolved trees are considered, there are still only 26/56 resolved nodes, but the TCA increases slightly to 0.397. This indicates that the low values are caused by genuine conflict between the loci, rather than by lack of resolution and systematic error,. The MRC groups species into the long-recognised subgenera, but relationships between and within these are often uncertain. The MCS clade is completely unresolved, except for the monophyly of the Venezuelan and Panamanian *H. hecale*. The *H. erato/sara* group is better resolved, although the precise relations between the major clades within this group are also not clear and the IC/ICA support for most species-level relations is unsatisfactory, indicating plausible alternative signals.

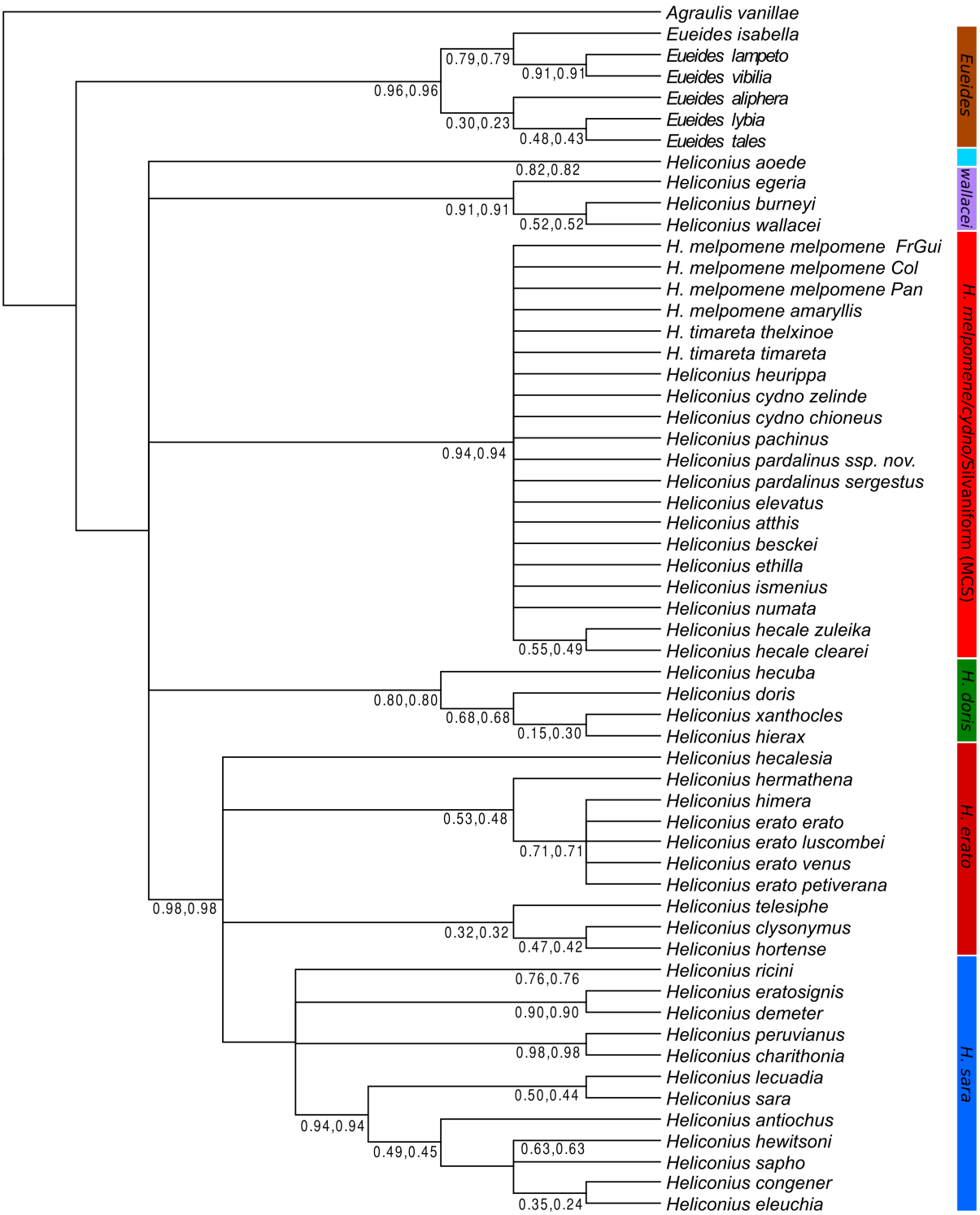

Agraulis vanillae
Eueides isabella
Eueides lampeto
Eueides vibilia
Eueides aliphera
Eueides lybia
Eueides tales
Heliconius aoede
Heliconius egeria
Heliconius burneyi
Heliconius wallacei
H. melpomene melpomene FrGui
H. melpomene melpomene Col
H. melpomene melpomene Pan
H. melpomene amaryllis
H. timareta thelxinoe
H. timareta timareta
Heliconius heurippa
Heliconius cydno zelinde
Heliconius cydno chioneus
Heliconius pachinus
Heliconius pardalinus ssp. nov.
Heliconius pardalinus sergestus
Heliconius elevatus
Heliconius atthis
Heliconius besckei
Heliconius ethilla
Heliconius ismenius
Heliconius numata
Heliconius hecale zuleika
Heliconius hecale clearei
Heliconius hecuba
Heliconius doris
Heliconius xanthocles
Heliconius hierax
Heliconius hecalesia
Heliconius hermathena
Heliconius himera
Heliconius erato erato
Heliconius erato luscombei
Heliconius erato venus
Heliconius erato petiverana
Heliconius telesiphe
Heliconius clysonymus
Heliconius hortense
Heliconius ricini
Heliconius eratosignis
Heliconius demeter
Heliconius peruvianus
Heliconius charithonia
Heliconius lecuadia
Heliconius sara
Heliconius antiochus
Heliconius hewitsoni
Heliconius sapho
Heliconius congener
Heliconius eleuchia
0.79,0.79
0.91,0.91
0.96,0.96
0.30,0.23
0.48,0.43
0.82,0.82
0.91,0.91
0.52,0.52
0.94,0.94
0.55,0.49
0.80,0.80
0.68,0.68
0.15,0.30
0.53,0.48
0.71,0.71
0.98,0.98
0.32,0.32
0.47,0.42
0.76,0.76
0.90,0.90
0.98,0.98
0.50,0.44
0.94,0.94
0.49,0.45
0.63,0.63
0.35,0.24
Eueides
wallacei
H. melpomene/cydno/Silvaniform (MCS)
H. doris
H. erato
H. sara

Figure 5.3. 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 a single, equally frequently observed 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.

Multispecies coalescent

The multispecies coalescent analysis produces a tree (Fig. 5.4) more resolved then the simple consensus (Fig. 5.3), but more ambiguous then the concatenation (Fig. 5.1). Most of the usual subgenera are recovered with confidence and separated by long branches. Exceptions include the relative placement of the "primitive" clades, position of the small *H. telesiphe* group and some species in the *H. sapho* clade, as well as two deep silvaniform nodes. Patterns of branch lengths in coalescent units are consistent with the chronogram (Fig. S5.2), as many of the nodes close in time are also weakly or not at all separated in the MP-EST tree due to incomplete sorting of individual lineages across the genome. Short internodal branches are also typically associated with poor support (black dots, Fig. 5.4).

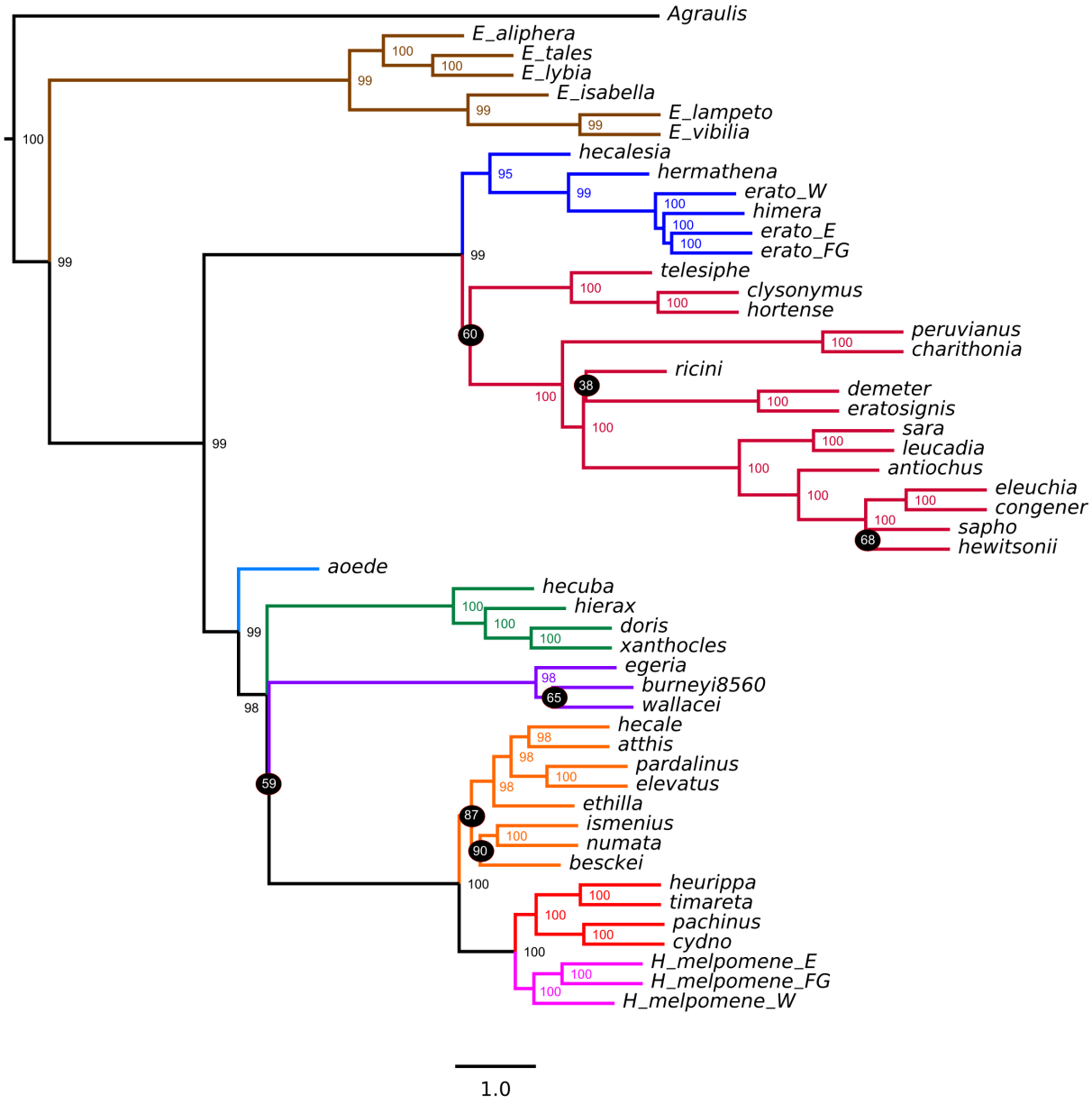


Figure 5.4. Incomplete lineage sorting at the autosomal loci. A multispecies coalescent tree estimated from the 6848 autosomal CDS gene trees under the MP-EST pseudolikelihood model shows lack of resolution at several nodes. Branch lengths in coalescent units, terminal branch lengths arbitrarily set to 1.0. Bootstrap support values indicated.

Genome-wide admixture

Clear results were obtained with the novel Ancestral Recombination Graph algorithm (Pickrell & Pritchard 2012). The inferred splits were roughly consistent with those found by a more precise ML method (Fig. 5.5) and also assigned relatively short branch lengths to the MCS clade. Intriguingly, the Magdalena Valley *H. melpomene* clustered closer to the *H. cydno* and cognates when not forced together with other samples in the West-of-Andes *H. melpomene* clade. The positive controls demonstrated that ARG performs well, as the results of Martin et al. (2013) were replicated by findings of high levels of gene flow ("migration weight") in sympatry between *H. timareta* and Eastern *H. melpomene* (Fig. 5.5: 1), as well as between Western *H. melpomene* and *H. cydno* (2). This is particularly reassuring, as the first comparison is based on a different selection of populations than in the original study. I also find evidence of moderate gene flow from Eastern *H. melpomene* into the sympatric *H. pardalinus/elevatus* lineage (3), as documented in a small sample of the genome by the Heliconius Genome Consortium (2012). An exciting novel finding is that limited gene flow occurred between the Amazonian *H. melpomene* and *H. numata* (4), as well as the crown group of silvaniforms (5). Bizarrely, the program always fitted a migration edge (7) between one of the deeper branches and any taxon set to be the outgroup (*H. aoede*, *Eueides* or *Agraulis*). This behaviour is likely to be an artefact of uncertainty in rooting, also reflected in the changing position of *H. aoede* across the analyses (Fig. 5.1-4, Sup. Fig. 5.1). Notably, no gene flow is inferred in *Eueides*, despite some instances of natural hybrids between multiple species (Mallet et al. 2007). The attempts to differentiate between hybridisation and ancestral polymorphism in reticulation networks based on gene trees failed due to excessive computational demands.

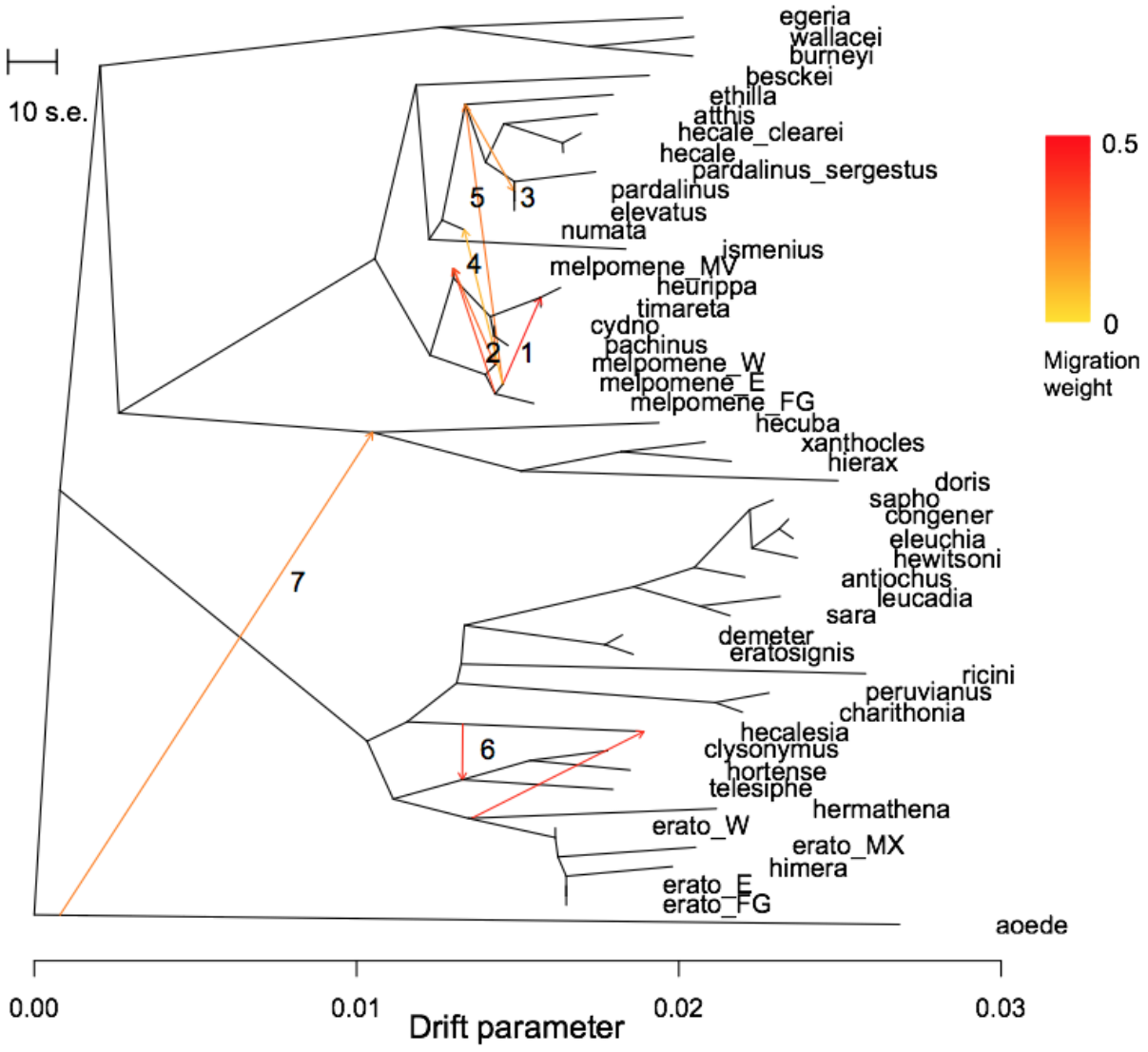


Figure 5.5. The extent of interspecific gene flow varies across the tree. TreeMix inference of splits and mixture from autosomal exonic SNPs. Migration edges (1-7) are inferred on a phylogenetic tree built from allele frequencies under a Gaussian genetic drift approximation. Colours of the edges correspond to the proportion of the genome exchanged.

An interesting problem is posed by *H. hecalesia* and the rest of the large *H. sapho/erato* clade. TreeMix places the species further from *H. erato* than any other analysis, but also infers high levels of hybridisation between lineages (Fig. 5.5: 7). The gene drift model coded into the TreeMix algorithm (Pickrell & Pritchard 2012) appears to confound migration with the signature of rapid lineage splitting, which would explain the estimates of nearly 50% gene flow around the *H. hecalesia* node (Fig. 5.5.6) and the nonsensical inference of gene flow between *Heliconius* and any arbitrary outgroup (Fig. 5.5.7). The hybridisation leading to *H. hecalesia* is further supported by the PCA plot of the first two principal components, which explain 50% of the variation, showing *H. hecalesia* as nearly equidistant from *H. erato*, *H. telesiphe* and *H. sapho* clades (Fig. 5.6). A STRUCTURE-like analysis (DAPC) found ten genotypic clusters, equivalent to the number of included species and *H. erato* lineages. All samples are assigned to the expected species and no admixture is inferred, perhaps as a result of the inherent bias of the Discriminant Analysis towards inferring between-population variability (Jombart et al. 2010). Nonetheless, further evidence of hybridisation is found at the colour pattern loci (see below).

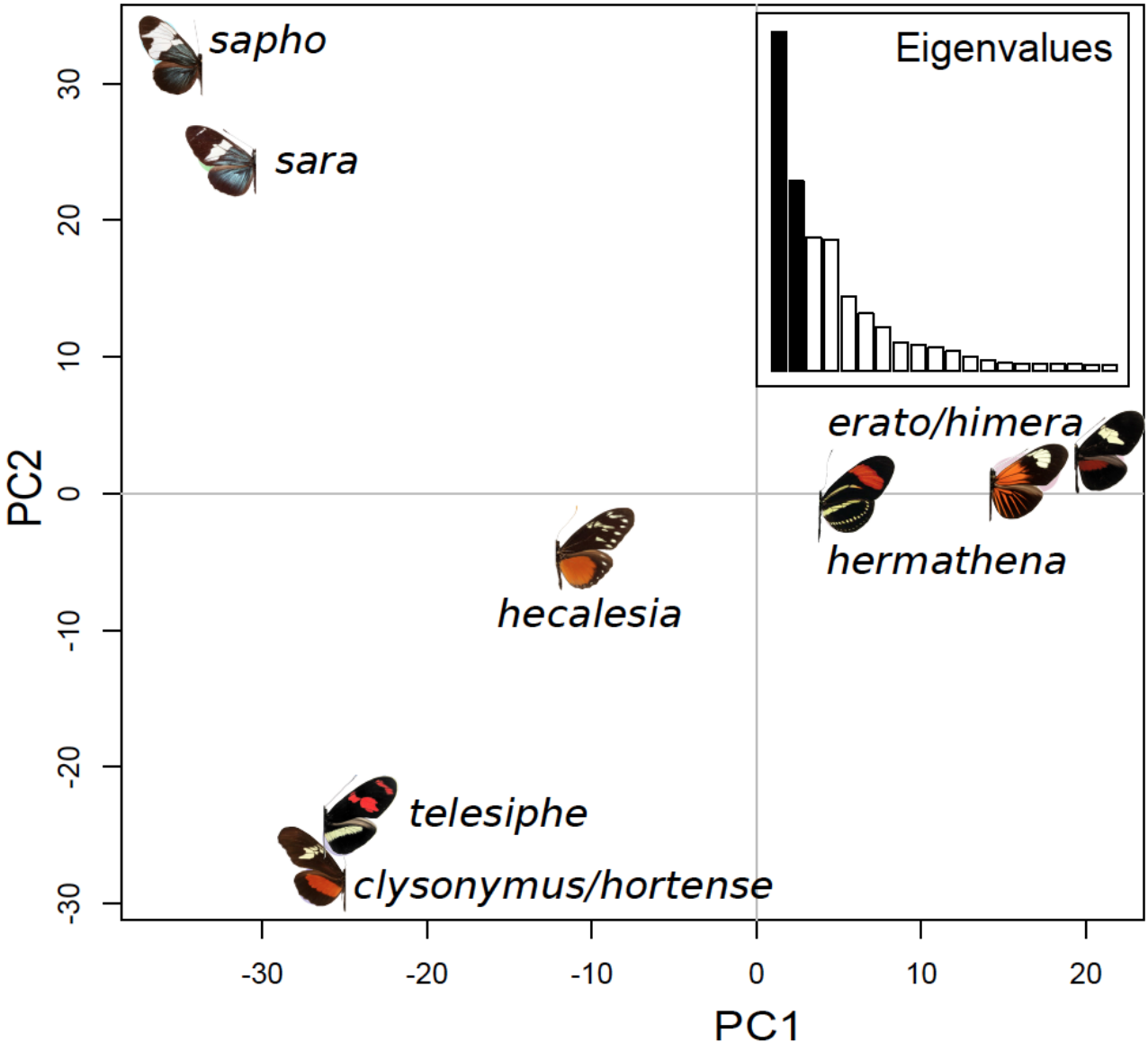


Figure 5.6. Ambiguous genomic composition of *Heliconius hecalesia*. Principal Component Analysis of variation in the autosomal exonic SNPs within the *H. erato/sapho* clade. First two PCs plotted, accounting for over half of the variation.

The Z chromosome

Unexpectedly, the large clade of silvaniforms is paraphyletic at the Z chromosome under both concatenation and coalescent (Fig. 5.10, S5.6). *Heliconius besckei/numata/ismenius* form a basal clade, sister to the group of all other silvaniforms and *H. melpomene/cydno*. Otherwise the Z chromosome shows slightly less gene tree incongruence (mean pairwise RF=0.19) and higher resolution than the autosomes (RF=0.23). The gene tree-species tree triplet distance is statistically significantly lower at the Z ($12.5*10^6$ vs $10.3*10^6$, Wilcoxon's test $p=3*10^{-11}$; Fig. S5.4). The Z-linked loci resolve 35/56 nodes in the 50% MRC tree, with a TCA (relative support) of 0.431, compared to the autosomal TCA of 0.322 (Fig. S5.5). Notably, many nodes within the MCS clade are resolved, and the *melpomene/cydno* clade appears monophyletic, with moderate support for *H. melpomene*. In fact, the amount of reticulation in a split network is higher in the *H. erato* clade than among the MCS. However, neither the autosomal nor the Z-linked consensus resolves the relative positions of the subgenera unequivocally. The concatenation tree is again very well resolved and supported. There is little support for the "faster X" hypothesis of faster accumulation of mutations at hemizygous chromosomes (Charlesworth et al. 1987). There is a statistically significant reduction in tree length of the Z as compared to the autosomes (Wilcoxon's test, $p<0.0001$), but the magnitude of the difference is small (Table 5.5) (Kayserili et al. 2012).

Introgression at the colour pattern loci

Gene trees constructed from 20 kbp sliding windows across the colour pattern scaffolds were variable, although large departures from the expected topology were infrequent. All of the regions, except 3/16 windows (40 kbp total) linked to the *K* locus (pattern shape), were completely informative and contained over 1000 good sites (average 19886 bp). In a few cases including four out of 60 *B/D* locus (red pattern) windows, a low

number of quality sites resulted in a surprising placement of *Eueides* within *Heliconius*. Due to the stochasticity in intraspecific topologies, the SH test was useless and found all the window trees to be significantly different from the species tree (MP-EST topology). In the following section I present the most striking and well supported results. Special attention is paid to the key regulatory and CDS regions (Table 5.2) (Nadeau et al. 2012; Supple et al. 2013; Nadeau et al. 2014; Wallbank et al. 2015; Reed et al. 2011).

At the *B/D* scaffold only four out of 60 windows show no departures from the species tree topology, but most of the variation is consistent with the genome-wide lack of resolution in the *H. melpomene/cydno*/silvaniform clade (MCS; Fig. 5.3). I found two novel introgressions. First, *H. hecalesia* frequently clusters with similarly patterned *H. clysonymus* and *H. hortense,* to the exclusion of their unquestionable relative in the species tree, *H. telesiphe* (Fig. 5.7). This relation is found exactly at the regions associated with red patterns in *H. erato* (Supple et al. 2013). A RelTime chronogram shows that *H. telesiphe* sequences diverged 2.9 MA (Fig. S5.2), whereas alleles of the other three species diverged much later, between 2.3-2.1 MA, which is also after speciation (Chapter 2). Second, alleles from *H. hecale clearei*, an unusual silvaniform with black and white patterns, cluster with *H. pardalinus/H. elevatus* sequences in eight windows and *H. numata* in one (Fig. 5.7).

All of the known introgressions are recapitulated, but some appear to involve larger regions than previously suggested. The association of *H. elevatus* and *H. melpomene/cydno* alleles (Heliconius Genome Consortium 2012) is observed in the five windows that include the enhancer regions for *Rays* and *Dennis* patterns and the *optix* CDS (Wallbank et al. 2015), but also in four windows upstream, three windows between the enhancers and *optix,* and three windows downstream. The greatest number of unusual clusters is seen at 360 to 380 kbp (Fig. 5.7), the section controlling both *H. melpomene* (Wallbank et al. 2015) and *H. erato* rays (Supple et al. 2013). Amazonian *H. melpomene*, *H. timareta contigua* (Ecuador) and

H. timareta florencia (Colombia) split off to form a clade sister to *H. cydno/pachinus*. However, *H. timareta thelxinoe* (Ecuador), *H. t. timareta* (Peru) and the southernmost Amazonian *H. melpomene amandus* are in their expected position with the Guianian *H. m. melpomene*. The alleles from *H. heurippa* form a clade within the Western *H. melpomene*, sister to the Colombian Magdalena Valley individual, supporting previous evidence for introgression of red patterns from *H. melpomene* into *H. heurippa* (Salazar et al., 2010).

The major locus *Yb* (called *Cr* in *H. erato*) encodes the yellow and white patterns (Nadeau et al. 2014; Sheppard et al. 1985). Mapping was compromised by high divergence (Nadeau et al. 2012), although none of the windows had insufficient data. Patterns at this locus are much less complex than at the *B/D*. Introgression from the eastern *H. melpomene* clade into *H. timareta* was found at four windows (310-330 kbp and 570-620 kbp from the start of the scaffold in the reference). The first region coincided with an F_{st} peak between *H. melpomene* and *H. timareta* (Nadeau et al. 2012) and clustered these species with silvaniforms. The second region lies where *H. cydno* and *H. timareta* sequences were similarly nested within the Western/Guianian *H. melpomene* (590-620 kbp), whereas *H. pardalinus/elevatus* was sister to *H. melpomene/cydno* (570-600 kbp). The latter association was also observed in the 670-690 kbp window containing the recently described *poikilomeusa* gene (Nadeau et al. 2014). Finally, I found a cluster of *H. hecalesia/hortense/clysonymus* alleles at 570-600 kbp.

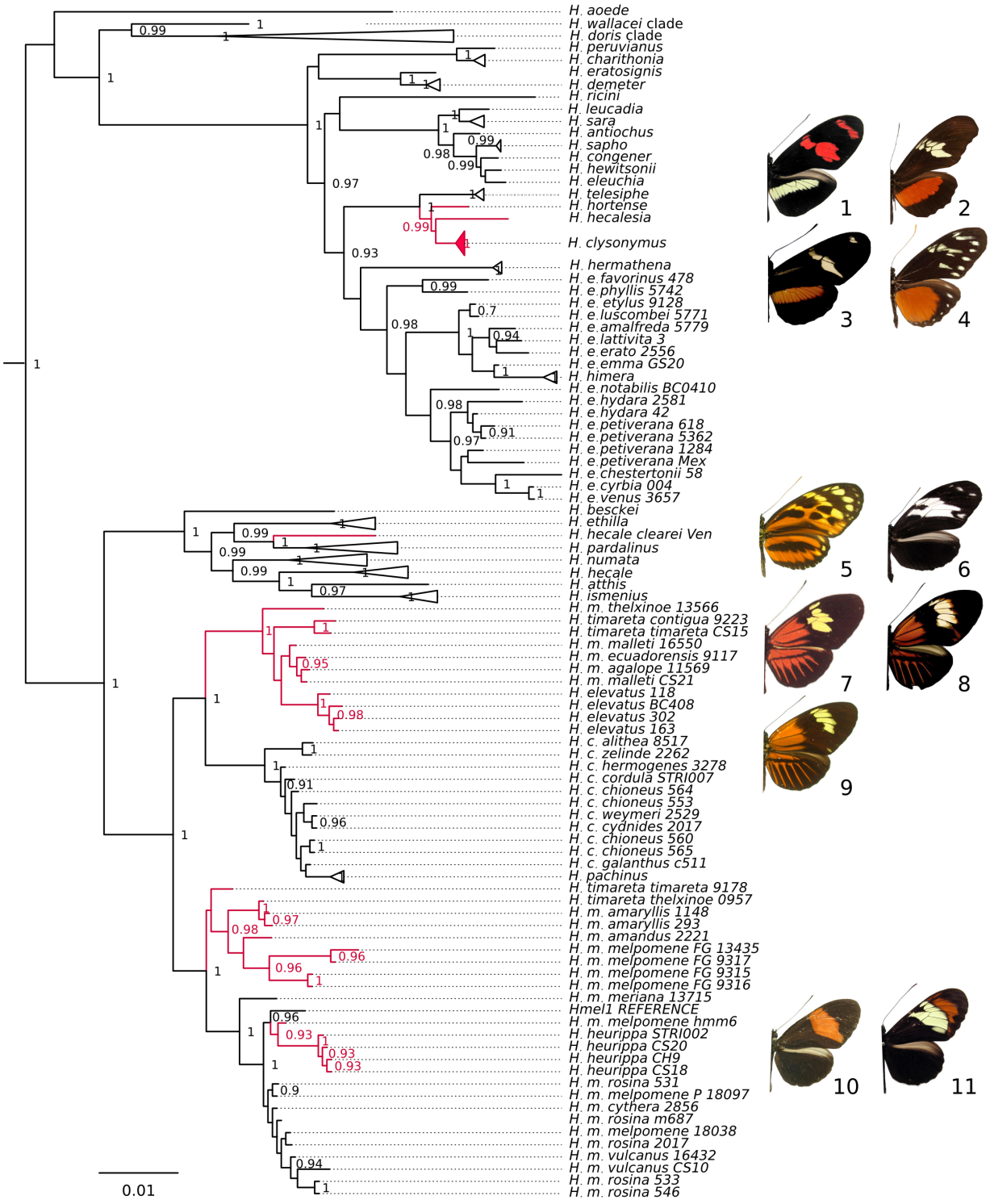

H. aoede
H. wallacei clade
H. doris clade
H. peruvianus
H. charithonia
H. eratosignis
H. demeter
H. ricini
H. leucadia
H. sara
H. antiochus
H. sapho
H. congener
H. hewitsonii
H. eleuchia
H. telesiphe
H. hortense
H. hecalesia
H. clysonymus
H. hermathena
H. e.favorinus 478
H. e.phyllis 5742
H. e. etylus 9128
H. e.luscombei 5771
H. e.amalfreda 5779
H. e.lattivita 3
H. e.erato 2556
H. e.emma GS20
H. himera
H. e.notabilis BC0410
H. e.hydara 2581
H. e.hydara 42
H. e.petiverana 618
H. e.petiverana 5362
H. e.petiverana 1284
H. e.petiverana Mex
H. e.chestertonii 58
H. e.cyrbia 004
H. e.venus 3657
H. besckei
H. ethilla
H. hecale clearei Ven
H. pardalinus
H. numata
H. hecale
H. atthis
H. ismenius
H. m. thelxinoe 13566
H. timareta contigua 9223
H. timareta timareta CS15
H. m. malleti 16550
H. m. ecuadorensis 9117
H. m. agalope 11569
H. m. malleti CS21
H. elevatus 118
H. elevatus BC408
H. elevatus 302
H. elevatus 163
H. c. alithea 8517
H. c. zelinde 2262
H. c. hermogenes 3278
H. c. cordula STRI007
H. c. chioneus 564
H. c. chioneus 553
H. c. weymeri 2529
H. c. cydnides 2017
H. c. chioneus 560
H. c. chioneus 565
H. c. galanthus c511
H. pachinus
H. timareta timareta 9178
H. timareta thelxinoe 0957
H. m. amaryllis 1148
H. m. amaryllis 293
H. m. amandus 2221
H. m. melpomene FG 13435
H. m. melpomene FG 9317
H. m. melpomene FG 9315
H. m. melpomene FG 9316
H. m. meriana 13715
Hmel1 REFERENCE
H. m. melpomene hmm6
H. heurippa STRI002
H. heurippa CS20
H. heurippa CH9
H. heurippa CS18
H. m. rosina 531
H. m. melpomene P 18097
H. m. cythera 2856
H. m. rosina m687
H. m. melpomene 18038
H. m. rosina 2017
H. m. vulcanus 16432
H. m. vulcanus CS10
H. m. rosina 533
H. m. rosina 546
0.01
1
2
3
4
5
6
7
8
9
10
11

Figure 5.7. Pervasive introgression across the species boundary at the red patterning locus. Branches in unexpected positions are 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*.

Two regions of topological variation were found at the forewing melanin shutter *Sd* (*Ac* in *H. erato*). In the 28-32 kbp region, which contains the *cuticular protein glycine-rich 24* (*Cpg24*) gene (Martin et al. 2012), *H. melpomene malleti* from the Eastern Cordillera of the Colombian Andes fell out at the base of *H. cydno/timareta*. In the region containing *WntA* and *ChitSynth* patterning genes (450-490 kbp) (Martin et al. 2012), I found evidence of the introgression events described above: the *H. hecalesia/hortense/clysonymus* cluster; *H. elevatus* basal to *H. melpomene*; *H. timareta* nested among eastern *H. melpomene*; *H. heurippa* alleles within *H. cydno*. The following scaffold in the linkage map contains the gene *defective proboscis extension response (dpr1*) (Martin et al. 2012). At this locus *H. timareta* alleles associate with *H. melpomene* (110-130kbp).

Shape of the distal edge of the forewing is controlled by the *Ro* locus: the 8 associated SNPs in *H. erato* lie on the scaffold HE671554 (Nadeau et al. 2014). In trees from the surrounding region (70-200 kbp), *H. timareta/heurippa* sequences are again mixed with the Amazonian *H. melpomene*. At 10-30 kbp I also find an *H. hecalesia/hortense/clysonymus*

clade.

Analysis of the shape locus *K* was hindered by incomplete mapping, undoubtedly related to the poor assembly of the two corresponding scaffolds (J. Davey, *pers. comm.*). The alignments from individual window produced low-resolution trees. However, an analysis of the entire scaffolds (HE671246, HE670889) (Nadeau et al. 2014) points to yet another case of introgression from Eastern *H. melpomene* into *H. timareta*. It must be noted that all of the above results are at best suggestive and in order to distinguish introgression from ILS, individual cases should be investigated with a coalescent method like IMa2 (see Pardo-Díaz et al. 2012 for an example).

DISCUSSION

Genome-wide hybridisation in Heliconius

Phylogenomics has massively expanded the potential of comparative molecular biology and been already applied widely to both ancient clades (e.g. birds, Zhang et al. 2014; mammals, Song et al. 2012) and recently diverged taxa (e.g. *Drosophila simulans*, Garrigan et al. 2012; great apes, Scally et al. 2012; populations of monarch butterflies, Zhang et al. 2014), but not yet at the intermediate level of large, recently-radiated genera. I present the first genome-wide study of a lepidopteran genus, assess the robustness of the species phylogeny and identify previously unknown introgression events. My work marks a return to studying *Heliconius* from the comparative perspective, which for over a decade has been – with a few exceptions (Mallet et al. 2007, Briscoe et al. 2013) – temporarily eclipsed by detailed studies of very recently diverged races and species (e.g. Martin et al. 2013, Supple et al. 2013).

Contrary to the predictions made on the basis of previous research and natural

occurrence of hybrid individuals, gene flow between species of *Heliconius* is common only in the MCS clade. Within this group, the Ancestral Recombination Graph demonstrates previously undocumented cases of introgression across the *Heliconius* species boundary, including <10% of the genome exchanged between *H. numata* and the Amazonian *H. melpomene* (Fig. 5.5.4) and ~20% exchange between *H. melpomene* and the ancestral *H. ethilla*/*H. pardalinus* lineage (Fig. 5.5.5). In addition, the ARG qualitatively and quantitatively recapitulates previous findings, including introgression into *H. heurippa/timareta* (Fig. 5.5.1; Salazar et al. 2008; 2010; Nadeau et al. 2013), extensive *H. melpomene/cydno* gene flow (Fig. 5.5.2; Martin et al. 2013; 2015), and exchanges between *H. melpomene* and *H. elevatus/pardalinus* (Fig. 5.5.3, 5.5.5; Heliconius Genome Consortium 2012; Wallbank et al. 2015; K. Dasmahapatra, unpublished). Intriguingly, the MCS clade is characterised by relatively short branches in the ML tree estimated from concatenated SNPs (Fig. 5.1), which may be an effect of persistent gene flow scrambling variation between lineages. Variation in branch lengths is observed in thousands of phylogenetic studies and is generally explained as resulting from variation in the rate of substitution. Introgression has not been typically considered as an explanation of this observation, but I propose that it may be a common cause of branch length variation.

The number and complexity of the identified events is consistent with a systematic inspection of collections by Mallet and colleagues (2007), who identified 161 hybrid specimens, including 72 within the *H. melpomene/cydno* clade, 11 among silvaniforms, and 10 between the two clades (Dasmahapatra et al. 2007). Indeed, a majority of wild-caught MCS hybrids can be recreated by crossing in controlled conditions (Mavarez et al. 2006) and although hybrid individuals do not typically exceed 0.1% of the surveyed populations in the Andes (Jiggins et al. 2001a), this proportion is known to reach 7% of *H. melpomene/cydno* individuals in at least one locality in Venezuela (J. Mavarez, unpublished). The relative

proportion of hybrid individuals remains low, perhaps due to female sterility (Naisbit et al. 2002), mate discrimination (Jiggins et al. 2001) and strong natural selection against hybrid phenotypes, but it is nonetheless sufficient to allow gene flow across the genome (Martin et al. 2013), and similarly small proportions have been found in *Anopheles* mosquitoes characterised by immense genome-wide gene flow (Fontaine et al. 2015).

Since *Heliconius* is considered notable for the number of species hybridising in the wild (Mallet 2005) and in captivity (Gilbert 2003), it is surprising that the genomic footprint of gene flow is restricted mostly to the MCS subclade, and common only among the five species in the *H. melpomene/cydno* group. This difference is partly explained by the great diversity of recently evolved species in this clade. However, the erato/sapho clade has a similar overall age and is even more speciose (Fig. S5.2), Furthermore, recent analysis has suggested that gene flow between *H. melpomene* and *H. cydno* started only after ~1 MA of separation, so is not solely a result of incomplete speciation (Martin et al. 2015). The porosity of the species barrier may therefore be a unique characteristic of the MCS clade for some other reason, consistent with the disproportionately high count of hybrid specimens.

In contrast, none of the genomic data presented here suggest any intraspecific gene flow in *Eueides*, and an order of magnitude fewer hybrids are known from the wild, even though many of the species are also numerous (Mallet et al. 2007). Nonetheless, all the hybrid *Eueides* were caught in Mexico, whereas the specimens analysed in this study came from Peru. Geographical sampling is generally a limiting factor in this study, as generation of sufficient data from a structured sampling of evenly-spaced populations remains very challenging. Specifically, species from the MCS clade are much better represented than any others (Appendix: Samples), and may therefore include more populations that hybridise partly because of greater sampling effort. The phenomenon of highly spatially variable gene flow in *Anopheles* (Lee et al. 2013) provides evidence that sequencing additional samples from

around the ranges of *Heliconius* may lead to further discoveries. Indeed, many events, for instance the occasional introgression of mitochondria and the *Ac* locus in the valleys of Colombia (Fig. S5.1; Salazar et al. 2008), appear to be restricted to very specific localities. Future studies should focus on increasing the number of represented populations, especially from Central America and the poorly sampled centre and East of the Amazon basin. This expansion can be carried out cost-effectively using capture approaches (e.g. Nadeau et al. 2012). Inclusion of further *Eueides* would also be useful for studies of introgression and to elucidate the adaptive reasons (if such exist) for the disparate number of species in the two sister genera.

Recipient	*B/D*	*Yb/Cr*	*Ac/Sd*	*Ro*	*K*	frequency	Autosomes-wide
				Donor			
H. hecalesia	*H. clysonymus/ hortense*	*H. clysonymus/ hortense*	*H. clysonymus/ hortense*	*H. clysonymus/ 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/ 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/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/timareta*			0.000	
H. pardalinus/elevatus/ hecale/atthis/ethilla						0.008	*H. melpomene* E

Table 5.5. 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. Frequency of the clusters counted among autosomal gene trees. Novel findings highlighted in yellow.

Introgression at the colour pattern loci

A large number of genomic studies of interspecific gene flow have found introgressions of relatively small haplotypes, 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). Nowhere is the importance of adaptive introgression as clear as in *Heliconius*, where strong natural selection has driven introgression of wing patterning alleles between *H. melpomene* and *H. elevatus*, MCS clade species diverged as long as 5 Myr (Heliconius Genome Consortium 2012; Wallbank et al. 2015). For the first time, I show (a) several cases of introgression at the *Yb* and *Ac* loci regulating the yellow/white patterns; and (b) evidence for introgression outside the MCS clade between *H. hecalesia* and *H. hortense/clysonymus* (Table 5.5). Previous reports have demonstrated a bewildering complexity of allele sharing across the *H. melpomene/cydno* group, and narrowed the introgressing region down to a section of the *B/D* scaffold containing putative *cis*-regulatory elements of the central patterning switch *optix* (Pardo-Díaz et al. 2012). My findings for the *B/D* locus are also consistent with these studies (Fig. 5.7) and the genome-wide patterns of gene flow (Fig. 5.5).

Whereas, in all other animals studied to date adaptive introgression has been localised to a single part of the genome, *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 5.5). For instance, *H. elevatus* derives its *Ac* genes (Table 5.5) and parts of its *B/D* haplotypes (Wallbank et al. 2015) from the Amazonian *H. melpomene*, but some of the *Ac*

haplotype appears to be transmitted from *H. cydno*, possibly *via H. melpomene*. Similarly, fragments of the *B/D*, *Yb* and *Sd* sequences in the silvaniform *H. elevatus* appear to be derived from either *H. melpomene* or *H. cydno*.

The increased amount of gene tree incongruence at the adaptive colour pattern loci (Fig. 5.8) strongly suggests that natural selection is the agent that promotes introgression, leading to significant departures from the species tree (Fig. 5.7). To some extent the observed patterns may also result from neutral gene flow between species. However, the adaptive explanation is supported not only by the unusually high topological incongruence, but also by the pattern of long distance linkage disequilibrium: changes from the expected topology across over 600 kbp of the *B/D* locus are especially pronounced exactly at the windows associated with regulatory variants in *H. erato* (Supple et al. 2013) and *H. melpomene* (Wallbank et al. 2015). Conversely, as *H. h. clearei* displays no red patterns and selection does not act, the *B/D* haplotypes of this species appear in the position expected from the species tree (Fig. 5.7).

It seems unlikely that the observed phylogenetic patterns resulted from molecular convergence due to selection on the same narrow "loci of repeated evolution" (Martin and Orgogozo 2013), rather than introgression across the species boundary. Conceptually, these alternatives resemble the dichotomy in the origins of adaptations *within* a species – either by novel mutation after, or from standing variation before the selective agent arises (Barrett and Schluter 2008; Welch and Jiggins 2014). In this case, the pool of standing genetic variation would include other *Heliconius* species, resembling the bacterial "supergenomes" - collections of genomic elements that circulate among diverse prokaryotic strains (Puigbò et al. 2014). However, repeated *de novo* mutation leading to convergence appears less plausible here, as the loci were partitioned into windows of 20 kbp containing thousands of variable sites, whereas the number of SNPs determining specific wing patterns may be less than

a hundred (Supple et al. 2013). Furthermore, the hypothesis of molecular convergence would predict clustering of convergent alleles from more distantly related taxa (e.g. the *H. doris* group with other rayed species), which is not observed.

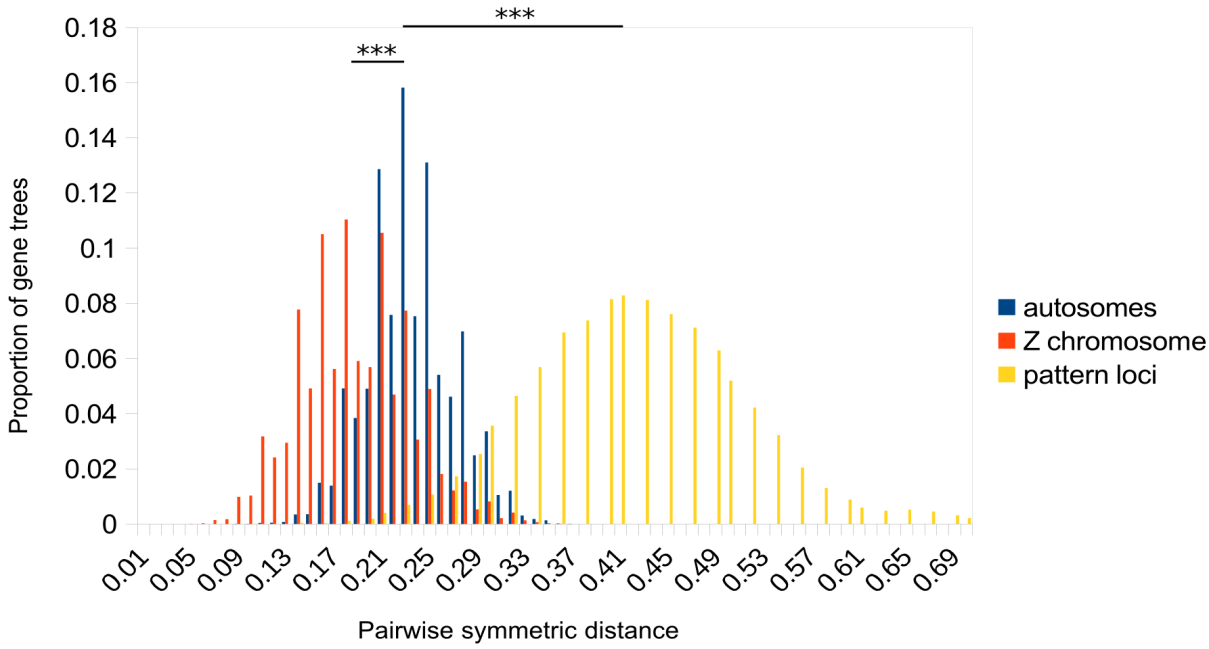


Figure 5.8. **Incongruence among gene trees is much higher at the colour pattern loci.** Distribution of a discordance measure (the symmetric pairwise Robinson-Foulds distance) among gene trees from the Z chromosome, the colour pattern loci, and other regions of the autosomes. Differences in means assessed with Wilcoxon's test (p<0.0001).

Nonetheless, the strategy of detecting adaptive introgression by inspection of gene trees in sliding windows has severe drawbacks that may be impacting our knowledge of systems ranging from yeast to mice (Heliconius Genome Consortium 2012; Martin et al. 2013; Supple et al. 2013; Fontaine et al. 2015; Yu et al. 2014; Lamichhaney et al. 2015).

In cases where the completeness of the alignments varies from window to window due to the changing quality of the mapping or assembly, many of the observed gene trees are unreliable due to limited variation, thus obscuring some patterns and making others appear unduly pronounced. Visual inspection of hundreds of phylogenies is prone to error and potentially subjective in the interpretation of what constitutes a sufficient deviation to warrant attention.

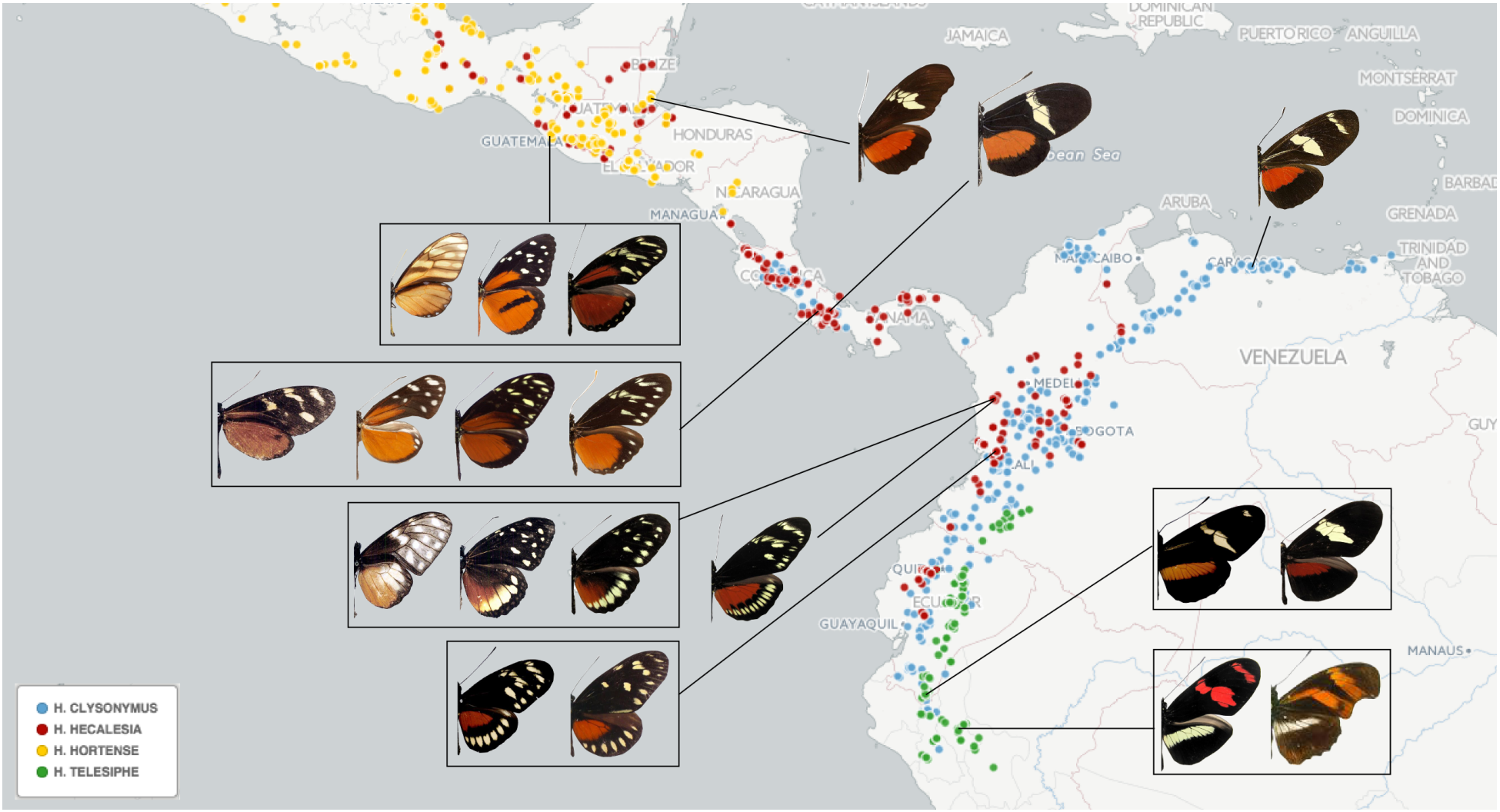
The repeated introgression at the colour pattern loci may also to an extent be a product of gene flow without selection, especially between *H. melpomene* and *H. cydno.* Even a naive calculation of frequency of the observed clusters among the CDS gene trees suggest that many of the patterns I found are common in the autosomal background (Table 5.5). Thus a clear way of testing the observed patterns against the null expectation is needed. Unfortunately, the coalescent Isolation-with-Migration modelling approach does not scale very well with the number of taxa or sites (Hey 2010). The *D* statistic, which dissects the signals of introgression and ILS, can be applied to a maximum of five taxa (Pease & Hahn 2014), and may be confounded by variation in sequence diversity when used to investigate small fragments of the genome in isolation (Martin et al. 2015a). My attempt to detect introgression in the entire genus is only a preliminary step and specific cases must be further confirmed with other techniques, for instance the linkage disequilibrium method used to demonstrate a selective sweep across the species barrier from *Drosophila simulans* into *D. sechelia* (Brand et al. 2013).

Genomic complexity in Heliconius hecalesia

The problem of distinguishing adaptive introgression from shared ancestral variation in *Heliconius* is exemplified by *H. hecalesia*, a relatively little-studied species with an unusual wide red bar on the hindwing (Fig. 5.9). At the key intervals within the *B/D* scaffold, as well as at *Cr, Sd* and *Ro, H. hecalesia* clusters with similarly patterned *H. clysonymus* and

H. hortense, to the exclusion of the phenotypically divergent *H. telesiphe*, suggesting an instance of interspecific allele sharing outside the MCS clade. However, the same topology is also observed at 14% of the autosomal gene trees. *Heliconius hecalesia* appears to have a highly mixed genomic composition (PCA, Fig. 5.6), but the uncertainty in phylogenetic signals appears to involve *hecalesia* and three clades making up the *H. erato* subgenus, as evidenced by the poor support for the *H. clysonymus* trio under concatenation (Fig. 5.1), disagreement between over half of the CDS trees (Fig. 5.3) and a notable reticulation in the split networks (Fig. 5.2). A similar pattern has been found in Darwin's finches (Lamichhaney et al. 2015), as well as at the base of mammals (Hallström & Janke 2010) and birds (Jarvis et al. 2014), and can be interpreted as a signature of explosive radiation during – respectively – habitat colonisation, climatic optimum or reduced competition. Two environmental factors may have facilitated a nearly simultaneous emergence of four *Heliconius* lineages that later diversified into three subgenera (*H. erato*, *H. sapho* and *H. clysonymus*) and *H. hecalesia*, resulting in a scrambled signal of unsorted ancient polymorphism. First, the groups vary considerably in their diets, with pronounced host specialisation typically on a single *Passiflora* species among the *H. sapho* group (Engler-Chaouat & Gilbert 2007) and a generalist tendency in the racially varied, widespread *H. erato*. However, this idea is weakly supported by empirical evidence, as the first two Principal Components show that host plant preference distinguishes between *H. hecalesia* and the *H. clysonymus* trio (Fig. S5.7), but not all 15 species. Another possibility is that the last phase of Andean orogeny in the North of the continent at around5 MA (Hoorn et al. 2010) increased elevation and latitudinal environmental gradients. The availability of new niches along a the slopes may have facilitated parapatric speciation, as *H. clysonymus* and *H. telesiphe* are mid-elevation specialists, while the other two species are found at lower elevations (Brown 1981; Brown & Benson 1975).

Figure 5.9. **Variation in patterning across the *Heliconius hecalesia* and *H. clysonymus* mimicry rings.** *Dircenna klugii* (female), *Tithorea tarracina duenna* and *Heliconius hecalesia octavia* (male); *Mechanitis polymnia isthmia*, *Hyposcada evanides*, *H. hecale zuleika*, *H. hecalesia formosus*; *Callithomia hezia tridactyla*, *Tithorea t. tarricina*, *H. h. hecalesia*; *H. h. ernestus*; *H. h. gunaesia*, *H. godmani*; *H. telesiphe*, *Podotricha telesiphe*; *H. clysonymus hygiana*, *H. himera*; *H. clysonymus montanus*; *H. ricini*. Figure copyrights: M. Demaio, G. Lamas, K. Davis, M. Stangeland, A. Warren; Godman and Salvin (1902). Map generated at www.cartodb.org from Rosser et al. (2012) data.


H. CLYSONYMUS
H. HECALESIA
H. HORTENSE
H. TELESIPHE
JAMAICA
DOMINICAN REPUBLIC
PUERTO RICO
ANGUILLA
MONTSERRAT
DOMINICA
BARBADOS
GRENADA
TRINIDAD AND TOBAGO
ARUBA
BELIZE
GUATEMALA
HONDURAS
EL SALVADOR
NICARAGUA
MANAGUA
COSTA RICA
PANAMA
MARACAIBO
CARACAS
VENEZUELA
MEDELLIN
BOGOTA
CALI
QUITO
ECUADOR
GUAYAQUIL
MANAUS
GUYANA

The adaptive introgression scenario fits with a number of observations. Singular *H. clysonymus* x *H. hecalesia* and *H. hortense* x *H. hecalesia* hybrids are known from the wild, and *H. hecalesia* is sympatric with the other two species throughout its range (Fig. 5.9), including at mid-elevation. It is believed that the subtle changes in the wing morphology of *H. hecalesia* are driven primarily by quasi-Batesian mimicry of the highly toxic Ithomiinae, including a rare case of sexually dimorphic mimicry of *Tithorea tarricina duenna* (by male *H. hecalesia octavia*) and *Dircenna klugii* (by females) (Brown and Benson 1975). Nonetheless, there appears to be an increasing trend in the width of the hindwing band away from the Equator, related to the patterns of variation in *H. clysonymus* and *H. hortense* (Fig. 5.9), perhaps as a result of localised introgression of modifier loci in sympatry. The dating of the putatively introgressed loci suggests the gene flow event (2.9 MA) occurred after the split between *H. hecalesia* and the *H. clysonymus* lineage (6.28-4.18 MA), although the dating may be unreliable if genome-wide gene flow scrambles variation.

Sex-linked loci

The paraphyly of silvaniform *Heliconius* at the Z chromosome (Fig. S5.5-5.7) is very unexpected. However, this may represent the true speciation history, as far as such history can be known. The silvaniforms paraphyly would require a major change in our understanding of the phylogenetic history of the clade. Although silvaniforms *sensu lato* share the tiger patterning, this character may be ancestral to the whole MCS and have been lost in *H. melpomene/cydno*. It is also possible that this phenotype has evolved multiple times through convergence between *Eueides*, silvaniforms and several sympatric Ithomiinae.

Sex-linked markers have been suggested as more reliable than autosomes, primarily due to faster coalescence at lower effective population size (¾ of the autosomal N_e at the Z

and ¼ at the W), and reduced likelihood of hybridisation as a result of Haldane's Rule (Zhang et al. 2013; Fontaine et al. 2015). 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. 2015) have confirmed deeper coalescence and lower levels of phylogenetic discordance at the X chromosomes, leading to the conclusion that the genuine signal of species history is obscured at autosomal loci. A comparison of gene tree discordance levels, split networks and coalescent trees between autosomes and the Z demonstrates that the sex chromosome is indeed subject to lower levels of gene flow. The MCS clade shows the least reticulation in the split network (Fig. S5.6), whereas the basal splits seem uncertain and the *H. erato* group fits the bifurcating tree model the least, thus reversing the trend found at the autosomes (Fig. 5.2). This pattern seems to result from reduced introgression at Z-linked loci in the MCS clade (Martin et al 2014), but a lack of a similar pattern in the *H. erato* clade. The reasons for this difference between the two clades remains unknown.

Reduced introgression at the Z chromosome is typically explained as resulting from a concentration of speciation genes on this chromosome. Most Lepidoptera show an “X-effect”, whereby the majority of genes causing sterility and inviability of interspecific hybrids are found at the Z chromosome (Prowell 1998). Crosses show that in *Heliconius* postzygotic isolation by hybrid inviability is associated with sex-linked loci (Jiggins et al. 2001; Naisbit et al. 2002). In contrast, premating isolation due to mate and microhabitat choice also plays a major role in speciation and is largely autosomal (Jiggins et al. 2001). In *H. melpomene* and *H. cydno* crosses have shown that mate preference is autosomal and linked to wing pattern loci, as is host plant preference (Merrill et al. 2011). Overall therefore, it remains unclear whether the Z chromosome has a disproportionate role in speciation in *Heliconius*, as has been proposed in other Lepidoptera. Interestingly, my results also shed some light on the reasons for Haldane's Rule in *Heliconius*. The comparison of substitution

rates between the Z and autosomes, previously applied to *Drosophila* (Kayserili et al. 2012), does not support the "faster X" hypothesis. It was previously proposed that dominance is the main cause for the Z-effect in *Heliconius* sterility (Jiggins et al. 2001; Naisbit et al. 2002), but this is the first time that this hypothesis has been evaluated at the scale of the entire chromosome.

The attempts at identifying female-specific, sex chromosome W-linked markers have failed here and in other studies (A. Pinharanda, unpublished). A large proportion of the genes maintained at sex chromosomes play a role in postzygotic isolation, which was recently suggested as their key role (Lima 2014). Therefore the W chromosome could be completely absent from *Heliconius*, as it is in many other Nymphalidae with a ZZ/Z0 sex determination system (Sahara et al. 2012). Alternatively, the putative W markers used here may be absent from the *Heliconius* chromosome due to high gene turnover, or the W may consist solely of repeat elements. In the latter case it may need to be physically dissected for sequencing, as in the Pyralidae moths (Traut et al. 2013) or the glanville fritillary (Ahola et al. 2014).

It is clear from the short branches in the coalescent tree (Fig. S5.5) that even the Z chromosome has retained enough ancient polymorphism to make the estimates of *H. erato* clade and *Eueides* phylogeny partially unreliable, which explains the poor resolution of these clades in Chapter 2. The whole mitochondrial data are a substitute for W-linked matriline markers, but do not show overwhelmingly higher support or length at the problematic branches, such as the divergences of *Neruda* and the *H. wallacei* clade (Fig. S5.1). Generally, the same nodes tend to be problematic in concatenation and coalescent analyses of mitochondrion-, Z- and autosome-linked markers (Fig. 5.4), even if no evidence of gene flow was found. This consistency of inconsistencies supports the idea that some of the diversification events among Heliconiini occurred nearly simultaneously and are unlikely to be ever satisfactorily "resolved" in a binary fashion.

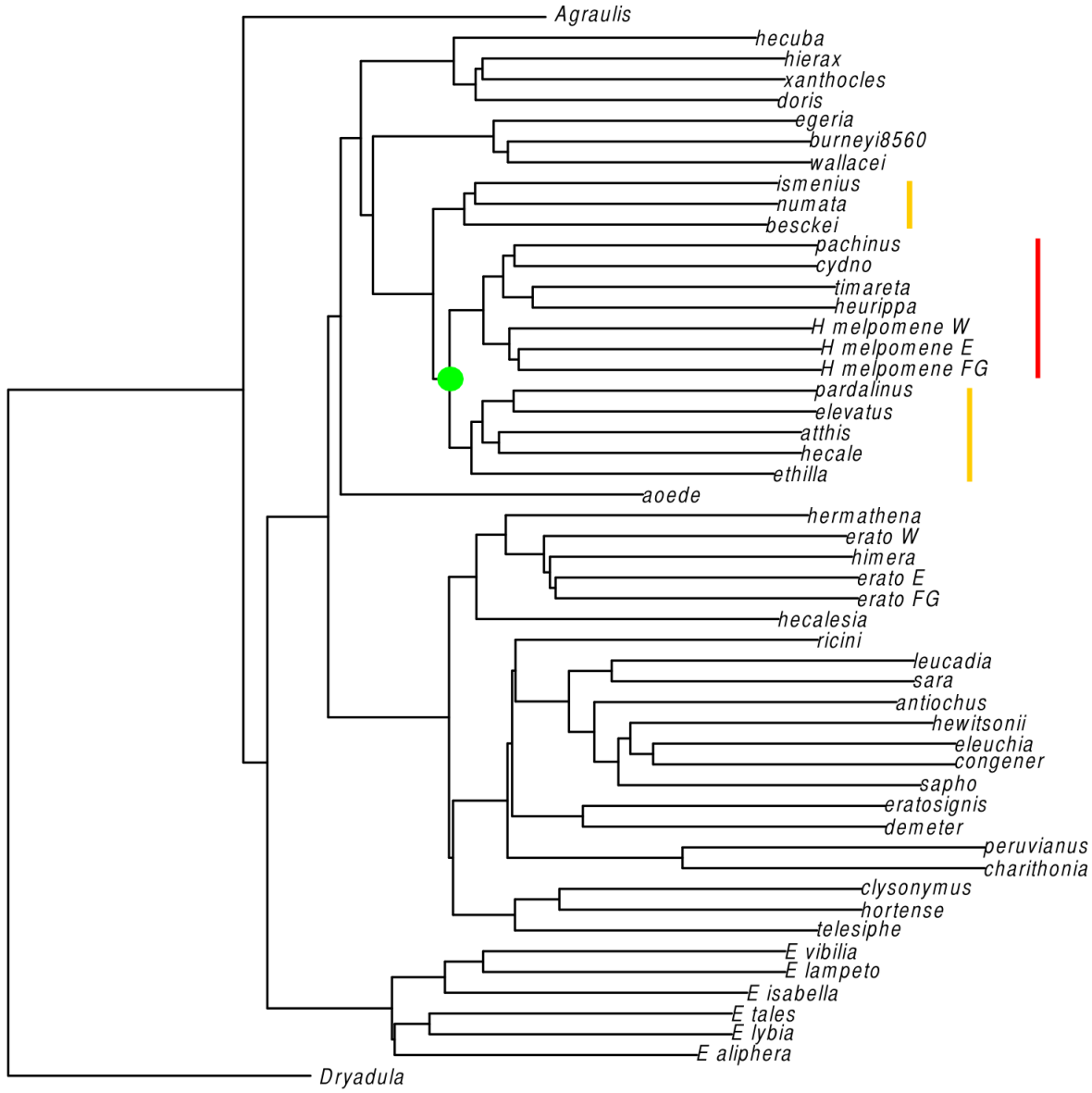


Figure S5.10. Silvaniforms are paraphyletic at the Z chromosome. A coalescent (MP-EST) tree based on 406 sex-linked gene trees shows unexpected shared ancestry (green node) linking the MCS clade (red bar) with a subset of silvaniforms (orange). Lengths of the branches are in coalescent units and thus meaningless for the terminal branches.

Technical improvements and implications for lepidopteran genomics

This study is a compromise between a range of choices regarding sample selection, genotyping and analytics. Although not all of these could be optimised within the constraints of time and resources, some obvious future improvements need to be reviewed. The selection of samples is perhaps the most comprehensive sampling yet attempted of an adaptive radiation, but is characterised by an excessive skew towards the most intensely studied clades – *H. melpomene*/silvaniforms and *H. erato/himera*. The lower representation of the smaller subclades may have exacerbated the taxon sampling error and made it more difficult to resolve the problematic nodes at the base of the tree and in the *H. sapho* group. In addition to deeper intraspecific sampling, studies of speciation and molecular adaptation will benefit greatly from the inclusion of further outgroups, as has been the case in the recent studies of bird (Zhang et al. 2014) and cichlid diversity (Brawand et al. 2014). Genomes of other Heliconiini genera and Acraeini will help to identify the factors responsible for relative advancement of *Heliconius*, including sensory and physiological innovations (Brown 1981).

The reference-mapping approach, although widely applied (e.g. monarch butterflies, Zhang et al. 2014; finches, Lamichhaney et al. 2015) is limited by the inverse relation between evolutionary divergence and the quality of mapping to the *H. melpomene* genome (Table 5.2). The simplest alternative is to assemble individual genomes *de novo* and align the contigs (the approach taken in comparisons of avian genomes, Zhang et al. 2014). An obvious advantage of this technique would be the recovery of greater amounts of non-coding sequence, which could be identified and homologised between closely related species to extend the dataset beyond the exome. The disadvantages include immense computational power required to identify homologous loci between all the possible combinations of 145 samples, greater potential to collapse the reads from the paralogues into chimeric contigs (Simpson et al. 2009), and possible variation in assembly quality resulting from differences in

the depth of sequencing. An intermediate step between the two extremes would be either to use multiple draft references for specific subgenera (e.g. *H. erato*; W. O. McMillan, unpublished), as recently done for *Anopheles* (Fontaine et al. 2015), or to carry out a reference-guided assembly, as notably done for the grouse genome (Wang et al. 2014). Yet both of these options leave the difficulty of homology-assignment unresolved.

The biggest constraint on the reference approach is the moderate quality of the *Hmel1* draft assembly (Ahola et al. 2014), which consists of 4309 individual scaffolds (N50=21 kbp), including only ~1500 of length sufficient to include in sliding-windows analyses (Martin et al. 2013), as compared to full chromosome assemblies available for other insect radiations (Clark et al. 2007; Neafsey et al. 2014). Further 67 Mbp of sequence are contained in redundant haplotype scaffolds (Heliconius Genome Consortium 2012), which were not included in my experiments. Unfortunately, these contain important sequences: a BLAST search reveals that V*ermillion* and S*carlet*, two out of six genes in the ommochrome pigmentation pathway (Ferguson and Jiggins 2009), are currently found only in the haplotype set not included in the reference. Contextualised characterisation of heterogeneity in the phylogenetic signal across the genome, leading to the detection of regions subject to introgression and selection (Pease and Hahn 2013), will require an improved draft scaffolded with RAD-seq data and third generation long reads (J. Davey, unpublished). This will enable identification of regions characterised by low recombination, least likely to be subject to ILS (Pease & Hahn 2013).

The genotyping pipeline used in this work is not substantially different from previous studies (Martin et al. 2013, Supple et al. 2013), but testing the tools and determining optimal settings was necessary to ensure the quality of the downstream analysis (Davey 2013). In a recent study, Greminger and colleagues (2014) demonstrated that three different genotyping algorithms disagree on up to 43% of all SNP calls when applied to high coverage primate data. The confidence in individual variant calls is arguably not critical in case of

phylogenomic studies, as random genotyping errors are unlikely to bias the signal in any specific way. However, I would caution against using the data from my work or similarly called genotypes to study SNPs associated with specific phenotypes, until quality scores can be confidently recalibrated against carefully curated panels of known polymorphisms, such as recently published by Keightley et al. (2014).

The difficulties in phylogenetic reconstruction outlined in previous chapters have been addressed by combining distinct methods, but further advances could be made with non-coding sequence data. Resolution of the avian family tree increases markedly when intronic sequences with higher substitution rates are used instead or in addition to the exome data (Jarvis et al. 2014). However, no benefit of using fast evolving single-copy genes was found in fungi (Aguileta et al. 2008) or in simulations of more relevant timescales. A small number of highly variable loci appears sufficient to markedly improve the estimates (Lanier et al. 2014). This helps to explain why the overall support (e.g. the TCA; Salichos & Rokas 2013) does not increase when only the gene trees from the 1000 longest alignments are used, despite the expectation that unresolved phylogenies will reduce the efficacy of MSC algorithms (Lanier et al. 2014; Mirarab et al. 2014). However, more confidence could be placed in a tree inferred with a method that bins gene trees by similarity (Mirarab et al. 2014), especially if the bins could be defined based on linkage disequilibrium (Ané 2011). The number of variants available for the supermatrix analysis was certainly sufficient, but the detected compositional heterogeneity (X^2-test in PAUP*, $p<0.00001$) could lead to compositional attraction. This problem is not observed, as the application of the LogDet model (Fig. S5.2) (Steel 2005) did not change the topology substantially. All of the above problems could be further addressed by focusing on molecular morphology characters like presence and absence of retrotransposons, which show relatively low homoplasy (Jarvis et al. 2014). Technically this would be a difficult task, as transposable elements constitute 25% of the *H. melpomene*

genome (Lavoie et al. 2013), and assignment of short read homology to correct reference loci may be unreliable.

Conclusion

The exome-based analysis of 90% of the *Heliconius* species reveals high levels of incongruence across the genomes and identifies the predominant reason as incomplete lineage sorting resulting from rapid divergences. Gene flow between species is limited to the MCS clade and occurs at especially high levels in the *H. melpomene* group, although *H. hecalesia* may constitute one case outside the MCS. Evidence is found for elevated levels of gene flow at all of the colour pattern loci, especially *B/D*, and for the first time *Yb/Cr* and *Ac/Sd*. Conversely, introgression is significantly limited at the Z sex chromosome, which probably reveals the most reliable phylogeny of the genus and does not support the monophyly of silvaniform butterflies. Autosomal, Z-linked and mitochondrial markers disagree widely on the order of speciation events. Contrary to the prevalent naïve expectation, increasing the amount of data without accompanying analytical sophistication does not solve the problem of finding the species tree of a rapid radiation, but leads to inflated confidence in incorrect results.

The study of adaptive radiations has been central to the development of modern evolutionary biology and provided the crucial comparative cases for studies of adaptation and speciation (Schluter 2000; Gavrilets and Losos 2009), both at the recent and increasingly at the deep timescales (dos Reis et al. 2014; Moen and Morlon 2014). Unfortunately, efforts to understand adaptive radiations presents a substantial and circular challenge, as detailed phylogenetic understanding is of paramount importance, but inherently complicated by rapid divergence and introgression – often the key properties of recent adaptive radiations (Glor 2010). Developing a macroevolutionary view of a radiation is therefore contingent upon our ability to resolve the history of speciation, in many cases taking into account hybridisation (Fontaine et al. 2015). These tasks have been recently transformed with the emergence of high-throughput sequencing data.

The rich *Heliconius* literature, which has inspired advances in our understanding of speciation (Abbott et al. 2013; Schumer et al. 2014; Seehausen et al. 2014) and adaptation (Hedrick 2013), has largely shunned macroevolutionary approaches, and only a few modern studies consider the entire radiation (e.g. Beltrán et al. 2007; Engler-Chaouat & Gilbert 2007; Rosser et al. 2012). My research combines a detailed analysis of genetic and genomic data to reconstruct phylogeny in the complex, recent radiation of the Heliconiini butterflies; an assessment of the sources of the phylogenetic incongruence; and finally testing of specific hypotheses on the drivers and dynamics of this charismatic radiation. The first three projects – focused respectively on incongruence and diversification, coevolution, and hybridisation – culminate in a pioneering phylogenomic study of the genus *Heliconius*. Taken

together, the individual chapters paint a comprehensive picture of both the pattern and the process leading to the diversity of a charismatic Neotropical taxon.

Chapter 2 introduces the problems inherent in the inference of a phylogeny and addresses the challenge with a combination of experimental approaches that emphasises identification and resolution of phylogenetic conflicts. Nearly taxonomically complete data make it possible to provide a detailed timeframe for future studies and have already been used to calibrate the timing of divergence with gene flow between two species of Heliconius (Martin et al. 2015) and the split timing between pairs of sister species (Rosser et al. 2015). Furthermore, I analysed the diversification rate (Fig. 2.1), inferring a plausible relation with concurrent geoclimatic changes in the Neotropics. Further evidence of environmental shifts driving allopatric and parapatric speciation is found in Chapter 3 (in *Passiflora*; Fig. 3.3) and Chapter 4 (in *Heliconius*; Fig. 4.6). In Chapter 3 I consider what could be the most prominent ecological factor affecting Heliconiini – their host plants – but surprisingly find evidence only for diffuse coevolution, which led to morphological adaptation rather than cospeciation. However, my novel chronogram of Passifloraceae (Fig. 3.2) suggests that the two groups may have diversified contemporaneously (Fig. 3.3).

In Chapters 4 and 5 I develop a pipeline to analyse genome-wide data from relatively distantly related species and use it to characterise the occurrence and extent of introgression between distinct taxa. *Heliconius* are widely cited as the prime example of the role that introgression plays in evolution, as both a formative agent of speciation (Jiggins et al. 2008; Abbott et al. 2013; Schumer et al. 2014) and a source of adaptive variation (Hedrick 2013). By adopting a genus-scale perspective I demonstrate that this property is in fact restricted to a subclade that comprises only a quarter of the heliconian diversity, showing that even in this radiation hybridisation plays a prominent, but phylogenetically restricted role.

Some of the projects included in this thesis reached led to negative conclusions, such as the absence of hybridisation in *H. hermathena* and the broader *H. erato* clade, or found equivocal results, such as diffuse coevolution between Heliconiini and the passion vines. However, my findings further the understanding of heliconian evolution not only by identifying the significant signals, but also precisely by highlighting what processes had limited or mixed effects on speciation, and how their importance differs between subclades. Inevitably, any conclusions drawn from the study of this radiation are limited in multiple ways. First, as pointed out in a recent review of heliconian biology, even if some clades are particularly useful in dissecting specific biological mechanisms, no single group can ever genuinely serve as a "model" in ecology and evolution (Merrill et al. 2015). Indeed, my work suggests that even within the Heliconiini radiation the emphasis on a single subclade may have exaggerated the importance of introgression (Chapter 5). Second, although with over 70 species heliconians are not a minor group, and comparisons can be made between several pairs of sister species (e.g. Rosser et al. 2015), comparative methods may not achieve suitable power unless hundreds of species are considered (FitzJohn 2010). Finally, many potentially key events took place only once and their importance is difficult to test (Chapter 2). While some authors argue that the problem of singular events is ultimately the doom of evolutionary biology and neutral explanations may be the only acceptable ones (Bokma et al. 2014), I object to this methodological fatalism. Although every species or clade investigated formed under unique circumstances, detailed studies of radiations add up to a picture challenging many of the earlier views on the nature of speciation in allopatry and without gene flow (Gavrilets and Losos 2009). Furthermore, the genomic perspective offers a new dimension, where generalisations glanced from radiation patterns can be tied to the underlying molecular mechanisms of change at the micro- and macroevolutionary scales.

Future research directions

In stark contrast to the better-explored but much less diverse vertebrates, among invertebrates comprehensive genus- and genome-wide data are available only for two fly genera (Clark et al. 2007; Fontaine et al. 2015). The data generated in my studies (exome surveys, genome-wide alignments, SNPs, gene alignments and trees) can be used productively to investigate many aspects of molecular evolution in lepidoptera and broadly at the level of a recent radiation [compare with the six cichlid genomes: (Brawand et al. 2014)], as well as to establish best practices in phylogenetics. Possible examples, some already under way, include:

- Discovery of loci under strong selection, and regions of high conservation (e.g. Conserved Noncoding Elements) that are likely to harbour transcription factor binding sites (J. Hanly, *pers. comm.*; Wallbank et al. 2015). This can be done easily using the TreeMix results (Pickrell 2012).
- Identification of introgressing loci and genes involved in the early stages of speciation, based on the relative depth of coalescence (Henning and Meyer 2014).
- Comprehensive analysis of molecular evolution at other known adaptive loci, for instance the ommochrome pathway gene *cinnabar* (K. Kozak, unpublished).
- Quantification of the amount of ancient variation preserved at the colour pattern loci across the genome and likely to confound the introgression analyses (Brawand et al. 2014).
- Detection of novel miRNAs, likely to play a role in morphological development (Surridge et al. 2011; Quah et al. 2015).
- Assessment of the dynamics, diversity, phylogenetic utility and genomic distribution of retrotransposable elements, which constitute ~25% of the *H. melpomene* genome,

are and may perform regulatory functions (Lavoie et al. 2013; Brawand et al. 2014).

- Identification of the Ultraconserved Elements and the most informative loci for phylogenetic analyses of Nymphalidae. Such a reference library would facilitate the development of capture approaches for butterflies, as it has in the case of Sanger sequencing (Wahlberg and Wheat 2008).

- Design of SNP chips for cost-effective genotyping of large numbers of individuals from any Heliconiini species, including material from museum collections (e.g. a bovine microarray, Decker et al. 2009).

- Comparison and assessment of the efficacy of automated species delimitation algorithms in face of hybridisation. Small *Heliconius* datasets have been used previously as test cases (e.g. Zhang et al. 2011), although the utility of sequence-based computational delimitation in this group is debatable (Chapter 2; Elias et al. 2008). Heterogeneity of taxon sampling and hybridisation levels in our data, in conjunction with detailed taxonomy and knowledge of reproductive barriers, make for an interesting scenario.

My studies point to several further avenues for *Heliconius* research. For instance, the clade of *H. erato/H. sapho* is relatively understudied, yet unique in its phylogenetically conservative diet (Chapter 3, Fig. 3.5) and pupal-mating reproductive biology (Beltrán et al. 2007). Future work ought to investigate how these traits contributed to the rapid speciation and low rate of hybridisation in this group (Chapter 5, Fig. 5.4-5), which are likely related to chemical signalling and chemosensation (Estrada et al. 2011). Patterns exposed in Chapter 3 demonstrate highly labile host plant preferences among Heliconiini (Fig. 3.5) and suggest flexibility in physiological mechanisms of toxin sequestration (Spencer 1988), which ought to

be investigated at the molecular level. Finally, in the realm of genomics, much work remains to be done to understand which parts of the genome are prone to intraspecific introgression and why (Fig. 5.8; Nadeau et al. 2012; Martin et al. 2013), how this process changes across the speciation continuum (Martin et al. 2015), and what features make it prominent in some species but absent in many others (Chapter 5).

As high-throughput sequencing becomes the methodological standard for studies of biodiversity, opportunities arise to test long-standing ecological and evolutionary hypotheses, but effective analytical strategies are necessary. My work is among the first to capture genome-level patterns of divergence and gene flow across an entire radiation, building strong links between ecology, patterns of diversification and genomic signals. By re-examining a well-studied radiation from the macroevolutionary perspective and with genomic tools, I demonstrate unexpected heterogeneity in the prevalence of introgression, uncover possibly irreconcilable incongruence and verify previous generalisations. Future studies in the genomic era will further uncover the causes of heterogeneous patterns and processes at the scale of the entire radiation.

BIBLIOGRAPHY

1000genomes.org, (2015). *1000 Genomes*. [online] Available at: http://1000genomes.org [Accessed 7 Jun. 2015].

Abbott R., Albach D., Ansell S., Arntzen J.W., Baird S.J.E., Bierne N., Boughman J., Brelsford A., Buerkle C.A., Buggs R., Butlin R.K., Dieckmann U., Eroukhmanoff F., Grill A., Cahan S.H., Hermansen J.S., Hewitt G., Hudson A.G., Jiggins C., Jones J., Keller B., Marczewski T., Mallet J., Martinez-Rodriguez P., Möst M., Mullen S., Nichols R., Nolte A.W., Parisod C., Pfennig K., Rice A.M., Ritchie M.G., Seifert B., Smadja C.M., Stelkens R., Szymura J.M., Väinölä R., Wolf J.B.W., Zinner D. 2013. Hybridization and speciation. J. Evol. Biol. 26:229–46.

Aguileta G., Marthey S., Chiapello H., Lebrun M.-H., Rodolphe F., Fournier E., Gendrault-Jacquemard A., Giraud T. 2008. Assessing the Performance of Single-Copy Genes for Recovering Robust Phylogenies. Syst. Biol. 57:613–627.

Aguileta G., de Vienne D.M., Ross O.N., Hood M.E., Giraud T., Petit E., Gabaldón T. 2014. High variability of mitochondrial gene order among fungi. Genome Biol. Evol. 6:451–65.

Ahola V., Lehtonen R., Somervuo P., Salmela L., Koskinen P., Rastas P., Välimäki N., Paulin L., Kvist J., Wahlberg N., Tanskanen J., Hornett E.A., Ferguson L.C., Luo S., Cao Z., de Jong M.A., Duplouy A., Smolander O.-P., Vogel H., McCoy R.C., Qian K., Chong W.S., Zhang Q., Ahmad F., Haukka J.K., Joshi A., Salojärvi J., Wheat C.W., Grosse-Wilde E., Hughes D., Katainen R., Pitkänen E., Ylinen J., Waterhouse R.M., Turunen M., Vähärautio A., Ojanen S.P., Schulman A.H., Taipale M., Lawson D., Ukkonen E., Mäkinen V., Goldsmith M.R., Holm L., Auvinen P., Frilander M.J., Hanski I. 2014. The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. Nat. Commun. 5:4737.

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19:716–723.

Alberici da Barbiano L., Gompert Z., Aspbury A.S., Gabor C.R., Nice C.C. 2013. Population genomics reveals a possible history of backcrossing and recombination in the gynogenetic fish Poecilia formosa. Proc. Natl. Acad. Sci. U. S. A. 110:13797–802.

Alford M.H. 2005. Systematic studies in Flacourtiaceae. PhD thesis, Cornell University.

Allen M.B., Armstrong H.A. 2008. Arabia–Eurasia collision and the forcing of mid-Cenozoic global cooling. Palaeogeogr. Palaeoclimatol. Palaeoecol. 265:52–58.

Althoff D.M., Segraves K.A., Johnson M.T.J. 2014. Testing for coevolutionary diversification: linking pattern with process. Trends Ecol. Evol. 29:82–9.

Anderson C.N.K., Liu L., Pearl D., Edwards S. V. 2012. Tangled trees: the challenge of inferring species trees from coalescent and noncoalescent genes. Methods Mol. Biol. 856:3–28.

Andrés J.A., Larson E.L., Bogdanowicz S.M., Harrison R.G. 2013. Patterns of transcriptome divergence in the male accessory gland of two closely related species of field crickets. Genetics. 193:501–13.

Andrews S. 2014. FastQC.

Ané C., Larget B., Baum D. a, Smith S.D., Rokas A. 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24:412–26.

Ané C. 2011. Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. Genome Biol. Evol. 3:246–58.

Anisimova M., Gascuel O. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst. Biol. 55:539–52.

Arias C.F., Muñoz A.G., Jiggins C.D., Mavárez J., Bermingham E., Linares M. 2008. A hybrid zone provides evidence for incipient ecological speciation in Heliconius butterflies. Mol. Ecol. 17:4699–712.

Arias C.F., Salazar C., Rosales C., Kronforst M.R., Linares M., Bermingham E., McMillan W.O. 2014. Phylogeography of Heliconius cydno and its closest relatives: disentangling their origin and diversification. Mol. Ecol. 23:4137–52.

Arnold M. 2004. Evolution through genetic exchange. Oxford, UK: Oxford University Press.

Van der Auwera G.A., Carneiro M.O., Hartl C., Poplin R., Angel G. del, Levy-Moonshine A., Jordan T., Shakir K., Roazen D., Thibault J., Banks E., Garimella K. V., Altshuler D., Gabriel S., DePristo M.A. 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr. Protoc. Bioinforma. 43:1–33.

Baack E.J., Rieseberg L.H. 2007. A genomic view of introgression and hybrid speciation. Curr. Opin. Genet. Dev. 17:513–8.

Bachtrog D., Thornton K., Clark A., Andolfatto P. 2006. Extensive introgression of mitochondrial DNA relative to nuclear genes in the Drosophila yakuba species group. Evolution. 60:292–302.

Baele G., Lemey P., Bedford T., Rambaut A., Suchard M.A., Alekseyenko A. V. 2012. Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty. Mol. Biol. Evol. 29:2157–67.

Baird S.J.E., Ribas A., Macholán M., Albrecht T., Piálek J., Goüy de Bellocq J. 2012. Where

are the wormy mice? A reexamination of hybrid parasitism in the European house mouse hybrid zone. Evolution. 66:2757–72.

Bapst D.W. 2012. paleotree : an R package for paleontological and phylogenetic analyses of evolution. Methods Ecol. Evol. 3:803–807.

Barrett R.D.H., Schluter D. 2008. Adaptation from standing genetic variation. Trends Ecol. Evol. 23:38–44.

Barrow L.N., Ralicki H.F., Emme S.A., Lemmon E.M. 2014. Species tree estimation of North American chorus frogs (Hylidae: Pseudacris) with parallel tagged amplicon sequencing. Mol. Phylogenet. Evol. 75:78–90.

Bates H.W. 1863. The Naturalist on the River Amazons. London: John Murray.

Baum D.A. 2007. Concordance trees , concordance factors , and the exploration of reticulate genealogy. Taxon. 56:417–426.

Baxter S.W., Johnston S.E., Jiggins C.D. 2009. Butterfly speciation and the distribution of gene effect sizes fixed during adaptation. Heredity (Edinb). 102:57–65.

Bazinet A.L., Cummings M.P., Mitter K.T., Mitter C.W. 2013. Can RNA-Seq resolve the rapid radiation of advanced moths and butterflies (Hexapoda: Lepidoptera: Apoditrysia)? An exploratory study. PLoS One. 8:e82615.

Beccaloni G.W., Viloria A.L., Hall S.K., Robinson G.S. 2008. Catalogue of the Hostplants of the Neotropical Butterflies/ Catálogo de las Plantas Huésped de las Mariposas Neotropicales. m3m-. m3m - Monografías Tercer Milenio, Volume 8. Zaragoza, Spain: Sociedad Entomológica Aragonesa (SEA)/Red Iberoamericana de Biogeografía.

Belfiore N.M., Liu L., Moritz C. 2008. Multilocus phylogenetics of a rapid radiation in the genus Thomomys (Rodentia: Geomyidae). Syst. Biol. 57:294–310.

Beltrán M., Jiggins C.D., Brower A.V.Z., Bermingham E., Mallet J. 2007. Do pollen feeding , pupal-mating and larval gregariousness have a single origin in Heliconius butterflies ? Inferences from multilocus DNA sequence data. Society.:221–239.

Beltrán M., Jiggins C.D., Bull V., Linares M., Mallet J., McMillan W.O., Bermingham E. 2002. Phylogenetic discordance at the species boundary: comparative gene genealogies among rapidly radiating Heliconius butterflies. Mol. Biol. Evol. 19:2176–90.

Benson W.W. 1972. Natural selection for Müllerian mimicry in heliconius erato in Costa Rica. Science (80-.). 176:936–939.

Benson W.W.., Brown K.S., Gilbert L.E.. G. 1975. Coevolution of Plants and Herbivores : Passion Flower Butterflies. Evolution (N. Y). 29:659–680.

Bergsten J., Nilsson A.N., Ronquist F. 2013. Bayesian tests of topology hypotheses with an example from diving beetles. Syst. Biol. 62:660–73.

Bernaud D. 2015. Le site des Acraea de Dominique Bernaud. Available from

www.acraea.com.

Blandin P., Purser B. 2013. Evolution and diversification of Neotropical butterflies: Insights from the biogeography and phylogeny of the genus Morpho Fabricius, 1807 (Nymphalidae: Morphinae), with a review of the geodynamics of South America. Trop. Lepid. Res. 23.

Blomberg S.P., Garland T., Ives A.R. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution (N. Y). 57:717.

Boggs C.L., Smiley J.T., Gilbert L.E. 1981. Patterns of pollen exploitation by Heliconius butterflies. Oecologia. 48:284–289.

Bokma F., Baek S.K., Minnhagen P. 2014. 50 years of inordinate fondness. Sys. Bio., 63(2):251-256.

Bouckaert R.R. 2010. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics. 26:1372–3.

Boussau B., Szöllosi G.J., Duret L., Gouy M., Tannier E., Daubin V. 2013. Genome-scale coestimation of species and gene trees. Genome Res. 23:323–30.

Brand C.L., Kingan S.B., Wu L., Garrigan D. 2013. A selective sweep across species boundaries in Drosophila. Mol. Biol. Evol. 30:2177–86.

Brawand D., Wagner C.E., Li Y.I., Malinsky M., Keller I., Fan S., Simakov O., Ng A.Y., Lim Z.W., Bezault E., Turner-Maier J., Johnson J., Alcazar R., Noh H.J., Russell P., Aken B., Alföldi J., Amemiya C., Azzouzi N., Baroiller J.-F., Barloy-Hubler F., Berlin A., Bloomquist R., Carleton K.L., Conte M.A., D'Cotta H., Eshel O., Gaffney L., Galibert F., Gante H.F., Gnerre S., Greuter L., Guyon R., Haddad N.S., Haerty W., Harris R.M., Hofmann H.A., Hourlier T., Hulata G., Jaffe D.B., Lara M., Lee A.P., MacCallum I., Mwaiko S., Nikaido M., Nishihara H., Ozouf-Costaz C., Penman D.J., Przybylski D., Rakotomanga M., Renn S.C.P., Ribeiro F.J., Ron M., Salzburger W., Sanchez-Pulido L., Santos M.E., Searle S., Sharpe T., Swofford R., Tan F.J., Williams L., Young S., Yin S., Okada N., Kocher T.D., Miska E.A., Lander E.S., Venkatesh B., Fernald R.D., Meyer A., Ponting C.P., Streelman J.T., Lindblad-Toh K., Seehausen O., Di Palma F. 2014. The genomic substrate for adaptive radiation in African cichlid fish. Nature. 513:375–381.

Breinholt J.W., Kawahara A.Y. 2013. Phylotranscriptomics: saturated third codon positions radically influence the estimation of trees based on next-gen data. Genome Biol. Evol. 5:2082–92.

Briscoe A.D., Muños A.M., Kozak K.M., Walters J.R., Yuan F., Jamie G.A., Martin S.H., Dasmahapatra K., Ferguson L.C., Mallet J., Jacquin-Joly, Emmanuelle Jiggins C.D. 2013. Female Behaviour Drives Expression and Evolution of Gustatory Receptors in Butterflies. PLoS Genet. 9:e1003620.

Brower a V. 1994a. Rapid morphological radiation and convergence among races of the

butterfly Heliconius erato inferred from patterns of mitochondrial DNA evolution. Proc. Natl. Acad. Sci. U. S. A. 91:6491–5.

Brower a. V.Z., Egan M.G. 1997. Cladistic analysis of Heliconius butterflies and relatives (Nymphalidae: Heliconiiti): a revised phylogenetic position for Eueides based on sequences from mtDNA and a nuclear gene. Proc. R. Soc. B Biol. Sci. 264:969–977.

Brower a. V.Z. 1994b. Phylogeny of Heliconius butterflies inferred from mitochondrial DNA sequences (Lepidoptera: Nymphalidae). Mol. Phylogenet. Evol. 3:159 –174.

Brower A.V.Z. 1996. Parallel race formation and the evolution of mimicry in Heliconius butterflies : a phylogenetic hypothesis from mitochondrial DNA sequences. Evolution (N. Y). 50:195–221.

Brower A.V.Z. 1997. The evolution of ecologically important characters in Heliconius butterflies (Lepidoptera : Nymphalidae): a cladistic review. Zool. J. Linn. Soc. 119:457–472.

Brown K., Sheppard P., Turner J. 1974. Quaternary Refugia in Tropical America: Evidence from Race Formation in Heliconius Butterflies. Proc. R. Soc. London. Ser. B, Biol. Sci. 187:369–378.

Brown K.S., Benson W.W. 1977. Evolution in Modern Amazonian Non-Forest Islands : Heliconius hermathena. Biotropica. 9:95–117.

Brown K.S. 1981. The Biology of Heliconius and Related Genera. Annu. Rev. Entomol. 26:427–457.

Brudno M., Do C.B., Cooper G.M., Kim M.F., Davydov E., Green E.D., Sidow A., Batzoglou S. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. 13:721–31.

Brumfield R.T., Edwards S. V. 2007. Evolution into and out of the Andes: a Bayesian analysis of historical diversification in Thamnophilus antshrikes. Evolution. 61:346–67.

Bull V., Beltrán M., Jiggins C.D., McMillan W.O., Bermingham E., Mallet J. 2006. Polyphyly and gene flow between non-sibling Heliconius species. BMC Biol. 4:11.

Burbrink F.T., Pyron R.A. 2011. The impact of gene-tree/species-tree discordance on diversification-rate estimation. Evolution. 65:1851–61.

Bybee S.M., Yuan F., Ramstetter M.D., Llorente-Bousquets J., Reed R.D., Osorio D., Briscoe A.D. 2012. UV photoreceptors and UV-yellow wing pigments in Heliconius butterflies allow a color signal to serve both mimicry and intraspecific communication. Am. Nat. 179:38–51.

Cahill J.A., Stirling I., Kistler L., Salamzade R., Ersmark E., Fulton T.L., Stiller M., Green R.E., Shapiro B. 2015. Genomic evidence of geographically widespread effect of gene flow from polar bears into brown bears. Mol. Ecol. 24:1205–17.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421.

Capella-Gutiérrez S., Silla-Martínez J.M., Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25:1972–3.

Cardoso M.Z., Gilbert L.E. 2013. Pollen feeding, resource allocation and the evolution of chemical defence in passion vine butterflies. J. Evol. Biol. 26:1254–60.

Cardoso M.Z. 2008. Herbivore handling of a Plants trichome: the case of Heliconius charithonia (L.) (Lepidoptera: Nymphalidae) and Passiflora lobata (Killip) Hutch. (Passifloraceae). Neotrop. Entomol. 37:247–52.

Cartodb.com, (2015). *CartoDB is the easiest way to map and analyze your location data — CartoDB*. [online] Available at: http://cartodb.com [Accessed 5 Nov. 2015].

Ceccarelli F.S., Crozier R.H. 2007. Dynamics of the evolution of Batesian mimicry: molecular phylogenetic analysis of ant-mimicking Myrmarachne (Araneae: Salticidae) species and their ant models. J. Evol. Biol. 20:286–95.

Chamberlain N.L., Hill R.I., Baxter S.W., Jiggins C.D., Kronforst M.R. 2011. Comparative population genetics of a mimicry locus among hybridizing Heliconius butterfly species. Heredity (Edinb). 107:200–4.

Charlesworth B., Coyne J.A., Barton N.H. 1987. The Relative Rates of Evolution of Sex Chromosomes and Autosomes. Am. Nat. 130:113–146.

Chauhan R., Jones R., Wilkinson P., Pauchet Y., Ffrench-Constant R.H. 2013. Cytochrome P450-encoding genes from the Heliconius genome as candidates for cyanogenesis. Insect Mol. Biol. 22:532–40.

Cheng J., Czypionka T., Nolte A.W. 2013. The genomics of incompatibility factors and sex determination in hybridizing species of Cottus (Pisces). Heredity (Edinb). 111:520–9.

Chomicki G., Renner S.S. 2015. Phylogenetics and molecular clocks reveal the repeated evolution of ant-plants after the late Miocene in Africa and the early Miocene in Australasia and the Neotropics. New Phytol.

Clark A.G., Eisen M.B., Smith D.R., Bergman C.M., Oliver B., Markow T. a, Kaufman T.C., Kellis M., Gelbart W., Iyer V.N., Pollard D. a, Sackton T.B., Larracuente A.M., Singh N.D., Abad J.P., Abt D.N., Adryan B., Aguade M., Akashi H., Anderson W.W., Aquadro C.F., Ardell D.H., Arguello R., Artieri C.G., Barbash D. a, Barker D., Barsanti P., Batterham P., Batzoglou S., Begun D., Bhutkar A., Blanco E., Bosak S. a, Bradley R.K., Brand A.D., Brent M.R., Brooks A.N., Brown R.H., Butlin R.K., Caggese C., Calvi B.R., Bernardo de Carvalho a, Caspi A., Castrezana S., Celniker S.E., Chang J.L., Chapple C., Chatterji S., Chinwalla A., Civetta A., Clifton S.W., Comeron J.M., Costello J.C., Coyne J. a, Daub J., David R.G., Delcher A.L., Delehaunty K., Do C.B., Ebling H., Edwards K., Eickbush T., Evans J.D., Filipski A., Findeiss S., Freyhult E., Fulton L., Fulton R.,

Garcia A.C.L., Gardiner A., Garfield D. a, Garvin B.E., Gibson G., Gilbert D., Gnerre S., Godfrey J., Good R., Gotea V., Gravely B., Greenberg A.J., Griffiths-Jones S., Gross S., Guigo R., Gustafson E. a, Haerty W., Hahn M.W., Halligan D.L., Halpern A.L., Halter G.M., Han M. V, Heger A., Hillier L., Hinrichs A.S., Holmes I., Hoskins R. a, Hubisz M.J., Hultmark D., Huntley M. a, Jaffe D.B., Jagadeeshan S., Jeck W.R., Johnson J., Jones C.D., Jordan W.C., Karpen G.H., Kataoka E., Keightley P.D., Kheradpour P., Kirkness E.F., Koerich L.B., Kristiansen K., Kudrna D., Kulathinal R.J., Kumar S., Kwok R., Lander E., Langley C.H., Lapoint R., Lazzaro B.P., Lee S.-J., Levesque L., Li R., Lin C.-F., Lin M.F., Lindblad-Toh K., Llopart A., Long M., Low L., Lozovsky E., Lu J., Luo M., Machado C. a, Makalowski W., Marzo M., Matsuda M., Matzkin L., McAllister B., McBride C.S., McKernan B., McKernan K., Mendez-Lago M., Minx P., Mollenhauer M.U., Montooth K., Mount S.M., Mu X., Myers E., Negre B., Newfeld S., Nielsen R., Noor M. a F., O'Grady P., Pachter L., Papaceit M., Parisi M.J., Parisi M., Parts L., Pedersen J.S., Pesole G., Phillippy A.M., Ponting C.P., Pop M., Porcelli D., Powell J.R., Prohaska S., Pruitt K., Puig M., Quesneville H., Ram K.R., Rand D., Rasmussen M.D., Reed L.K., Reenan R., Reily A., Remington K. a, Rieger T.T., Ritchie M.G., Robin C., Rogers Y.-H., Rohde C., Rozas J., Rubenfield M.J., Ruiz A., Russo S., Salzberg S.L., Sanchez-Gracia A., Saranga D.J., Sato H., Schaeffer S.W., Schatz M.C., Schlenke T., Schwartz R., Segarra C., Singh R.S., Sirot L., Sirota M., Sisneros N.B., Smith C.D., Smith T.F., Spieth J., Stage D.E., Stark A., Stephan W., Strausberg R.L., Strempel S., Sturgill D., Sutton G., Sutton G.G., Tao W., Teichmann S., Tobari Y.N., Tomimura Y., Tsolas J.M., Valente V.L.S., Venter E., Venter J.C., Vicario S., Vieira F.G., Vilella A.J., Villasante A., Walenz B., Wang J., Wasserman M., Watts T., Wilson D., Wilson R.K., Wing R. a, Wolfner M.F., Wong A., Wong G.K.-S., Wu C.-I., Wu G., Yamamoto D., Yang H.-P., Yang S.-P., Yorke J. a, Yoshida K., Zdobnov E., Zhang P., Zhang Y., Zimin A. V, Baldwin J., Abdouelleil A., Abdulkadir J., Abebe A., Abera B., Abreu J., Acer S.C., Aftuck L., Alexander A., An P., Anderson E., Anderson S., Arachi H., Azer M., Bachantsang P., Barry A., Bayul T., Berlin A., Bessette D., Bloom T., Blye J., Boguslavskiy L., Bonnet C., Boukhgalter B., Bourzgui I., Brown A., Cahill P., Channer S., Cheshatsang Y., Chuda L., Citroen M., Collymore A., Cooke P., Costello M., D'Aco K., Daza R., De Haan G., DeGray S., DeMaso C., Dhargay N., Dooley K., Dooley E., Doricent M., Dorje P., Dorjee K., Dupes A., Elong R., Falk J., Farina A., Faro S., Ferguson D., Fisher S., Foley C.D., Franke A., Friedrich D., Gadbois L., Gearin G., Gearin C.R., Giannoukos G., Goode T., Graham J., Grandbois E., Grewal S., Gyaltsen K., Hafez N., Hagos B., Hall J., Henson C., Hollinger A., Honan T., Huard M.D., Hughes L., Hurhula B., Husby M.E., Kamat A., Kanga B., Kashin S., Khazanovich D., Kisner P., Lance K., Lara M., Lee W., Lennon N., Letendre F., LeVine R., Lipovsky A., Liu X., Liu J., Liu S., Lokyitsang T., Lokyitsang Y., Lubonja R., Lui A., MacDonald P., Magnisalis V., Maru K., Matthews C., McCusker W., McDonough S., Mehta T., Meldrim J., Meneus L., Mihai O., Mihalev A., Mihova T., Mittelman R., Mlenga V., Montmayeur A., Mulrain L., Navidi A., Naylor J., Negash T., Nguyen T., Nguyen N., Nicol R., Norbu C., Norbu N., Novod N., O'Neill B., Osman S., Markiewicz E., Oyono O.L., Patti C., Phunkhang

P., Pierre F., Priest M., Raghuraman S., Rege F., Reyes R., Rise C., Rogov P., Ross K., Ryan E., Settipalli S., Shea T., Sherpa N., Shi L., Shih D., Sparrow T., Spaulding J., Stalker J., Stange-Thomann N., Stavropoulos S., Stone C., Strader C., Tesfaye S., Thomson T., Thoulutsang Y., Thoulutsang D., Topham K., Topping I., Tsamla T., Vassiliev H., Vo A., Wangchuk T., Wangdi T., Weiand M., Wilkinson J., Wilson A., Yadav S., Young G., Yu Q., Zembek L., Zhong D., Zimmer A., Zwirko Z., Alvarez P., Brockman W., Butler J., Chin C., Grabherr M., Kleber M., Mauceli E., MacCallum I. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 450:203–18.

Clarkson C.S., Weetman D., Essandoh J., Yawson A.E., Maslen G., Manske M., Field S.G., Webster M., Antão T., MacInnis B., Kwiatkowski D., Donnelly M.J. 2014. Adaptive introgression between Anopheles sibling species eliminates a major genomic island but not reproductive isolation. Nat. Commun. 5:4248.

Colinvaux P.A., De Oliveira P.E., Bush M.B. 2000. Amazonian and neotropical plant communities on glacial time-scales: The failure of the aridity and refuge hypotheses. Quat. Sci. Rev. 19:141–169.

Cong Q., Borek D., Otwinowski Z., Grishin N. V. 2015. Tiger Swallowtail Genome Reveals Mechanisms for Speciation and Caterpillar Chemical Defense. Cell Rep. 10:910–919.

Conow C., Fielder D., Ovadia Y., Libeskind-Hadas R. 2010. Jane: a new tool for the cophylogeny reconstruction problem. Algorithms Mol. Biol. 5:16.

Constantino L.M., Salazar J.A. 2010. A review of the Philaethria dido species complex (Lepidoptera : Nymphalidae : Heliconiinae) and description of three new sibling species from Colombia and. Zootaxa. 27:1 – 27.

Counterman B.A., Araujo-perez F., Hines H.M., Baxter S.W., Morrison C.M., Lindstrom D.P., Papa R., Ferguson L., Joron M., Richard H., Smith C.P., Nielsen D.M., Chen R., Jiggins C.D., Reed R.D., Halder G., Mallet J., Mcmillan W.O. 2010. Genomic Hotspots for Adaptation : The Population genetics of Mullerian Mimicry in Heliconius erato. 6.

Coyne J.A., Orr H.A. 2004. Speciation. Sinauer Associates.

Crawford J., Riehle M.M., Guelbeogo W.M., Gneme A., Sagnon N., Vernick K.D., Nielsen R., Lazzaro B.P. 2014. Reticulate speciation and adaptive introgression in the Anopheles gambiae species complex. bioRxiv.:009837.

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 Evol. Biol. 10:210.

Cruaud A., Rønsted N., Chantarasuwan B., Chou L.S., Clement W.L., Couloux A., Cousins B., Genson G., Harrison R.D., Hanson P.E., Hossaert-McKey M., Jabbour-Zahab R., Jousselin E., Kerdelhué C., Kjellberg F., Lopez-Vaamonde C., Peebles J., Peng Y.-Q., Pereira R.A.S., Schramm T., Ubaidillah R., van Noort S., Weiblen G.D., Yang D.-R., Yodpinyanee A., Libeskind-Hadas R., Cook J.M., Rasplus J.-Y., Savolainen V. 2012. An

extreme case of plant-insect codiversification: figs and fig-pollinating wasps. Syst. Biol. 61:1029–47.

Cui R., Schumer M., Kruesi K., Walter R., Andolfatto P., Rosenthal G.G. 2013. Phylogenomics reveals extensive reticulate evolution in Xiphophorus fishes. Evolution. 67:2166–79.

Cuthill J.H., Charleston M. 2012. Phylogenetic codivergence supports coevolution of mimetic Heliconius butterflies. PLoS One. 7:e36464.

Cutter A.D. 2013. Integrating phylogenetics, phylogeography and population genetics through genomes and evolutionary theory. Mol. Phylogenet. Evol. 69:1172–1185.

Czypionka T., Cheng J., Pozhitkov A., Nolte A.W. 2012. Transcriptome changes after genome-wide admixture in invasive sculpins (Cottus). Mol. Ecol. 21:4797–810.

Dasmahapatra K.K., Lamas G., Simpson F., Mallet J. 2010. The anatomy of a “suture zone” in Amazonian butterflies: a coalescent-based test for vicariant geographic divergence and speciation. Mol. Ecol.

Davis C.C., Webb C.O., Wurdack K.J., Jaramillo C.A., Donoghue M.J. 2005. Explosive radiation of Malpighiales supports a mid-cretaceous origin of modern tropical rain forests. Am. Nat. 165:E36–65.

Day J.J., Peart C.R., Brown K.J., Friel J.P., Bills R., Moritz T. 2013. Continental diversification of an African catfish radiation (Mochokidae: Synodontis). Syst. Biol. 62:351–65.

Degnan J.H., Rosenberg N. a. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.

DePristo M.A., Banks E., Poplin R., Garimella K. V, Maguire J.R., Hartl C., Philippakis A.A., del Angel G., Rivas M.A., Hanna M., McKenna A., Fennell T.J., Kernytsky A.M., Sivachenko A.Y., Cibulskis K., Gabriel S.B., Altshuler D., Daly M.J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–8.

DePristo M.A. 2014. http://gatkforums.broadinstitute.org/discussion/1186/best-practice-variant-detection-with-the-gatk-v4-for-release-2-0. Available from http://gatkforums.broadinstitute.org/discussion/1186/best-practice-variant-detection-with-the-gatk-v4-for-release-2-0.

Derryberry E.P., Claramunt S., Derryberry G., Chesser R.T., Cracraft J., Aleixo A., Pérez-Emán J., Remsen J. V, Brumfield R.T. 2011. Lineage diversification and morphological evolution in a large-scale continental radiation: the neotropical ovenbirds and woodcreepers (aves: Furnariidae). Evolution. 65:2973–86.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond A.J., Rambaut A., Xie W. 2010. BEAuti.

Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–73.

Duenez-Guzman E. a, Mavárez J., Vose M.D., Gavrilets S. 2009. Case studies and mathematical models of ecological speciation. 4. Hybrid speciation in butterflies in a jungle. Evolution. 63:2611–26.

Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S. a, Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sørensen M. V, Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 452:745–9.

Durand E.Y., Patterson N., Reich D., Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28:2239–52.

Eaton D.A.R., Ree R.H. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae). Syst. Biol. 62:689–706.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–7.

Edwards S. V, Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. U. S. A. 104:5936–41.

Edwards S. V. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63:1–19.

Ehrlich P., Raven P. 1964. Butterflies and plants: a study in coevolution. Evolution (N. Y). 18:586–608.

Elias M., Gompert Z., Jiggins C., Willmott K. 2008. Mutualistic interactions drive ecological niche convergence in a diverse butterfly community. PLoS Biol. 6:2642–9.

Elias M., Hill R.I., Willmott K.R., Dasmahapatra K.K., Brower A.V.Z., Mallet J., Jiggins C.D. 2007. Limited performance of DNA barcoding in a diverse community of tropical butterflies. Proc. Biol. Sci. 274:2881–9.

Emsley M.G. 1963. A Morphological Study of Imagine Heliconiinae (Lep. Nymphalidae With a Consideration of the Evolutionary Relationships Within the Group. Zoologica. 48:85–129.

Engler H.S., Spencer K.C., Gilbert L.E. 2000. Preventing cyanide release from leaves. Nature. 406:144–5.

Engler-Chaouat H.S., Gilbert L.E. 2007. De novo synthesis vs. sequestration: negatively correlated metabolic traits and the evolution of host plant specialization in cyanogenic butterflies. J. Chem. Ecol. 33:25–42.

Eol.org, (2015). *Encyclopedia of Life - Animals - Plants - Pictures & Information*. [online] Available at: http://www.eol.org [Accessed 7 Jun. 2015].

Eriksson A., Manica A. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc. Natl. Acad. Sci. U. S. A. 109:13956–60.

Ermini L., Der Sarkissian C., Willerslev E., Orlando L. 2014. Major transitions in human evolution revisited: A tribute to ancient DNA. J. Hum. Evol. 79:4–20.

Etienne R.S., Haegeman B., Stadler T., Aze T., Pearson P.N., Purvis A., Phillimore A.B., Pearson N. 2012. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. Proc. Biol. Sci. 279:1300–9.

Etienne R.S., Haegeman B. 2012. A conceptual and statistical framework for adaptive radiations with a key role for diversity dependence. Am. Nat. 180:E75–89.

Feder J.L., Flaxman S.M., Egan S.P., Comeault A.A., Nosil P. 2013. Geographic Mode of Speciation and Genomic Divergence. Annu. Rev. Ecol. Evol. Syst. 44:73–97.

Felsenstein J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution (N. Y). 39:783–791.

Felsenstein J. 2004. Inferring phylogenies. Sunderland, Mass: Sinauer Associates.

Fennell T. 2010. Picard Tools.

Ferguson L., Marlétaz F., Carter J.-M., Taylor W.R., Gibbs M., Breuker C.J., Holland P.W.H. 2014. Ancient expansion of the hox cluster in lepidoptera generated four homeobox genes implicated in extra-embryonic tissue formation. PLoS Genet. 10:e1004698.

Ferguson L.C., Jiggins C.D. 2009. Shared and divergent expression domains on mimetic Heliconius wings. Evol. Dev. 11:498–512.

Ferrer-Paris J.R., Sánchez-Mercado A., Viloria Á.L., Donaldson J. 2013. Congruence and diversity of butterfly-host plant associations at higher taxonomic levels. PLoS One. 8:e63570.

Feuillet C., MacDougal J. 2003. A new infrageneric classification of Passiflora. Passiflora. 13:34–38.

Filipski A., Murillo O., Freydenzon A., Tamura K., Kumar S. 2014. Prospects for Building Large Timetrees Using Molecular Data with Incomplete Gene Coverage among Species. Mol. Biol. Evol. 31:2542–2550.

FitzJohn, R.G. 2010. Quantitative traits and diversification. Sys. Bio. 59(6):619-633.

Flanagan N.S., Tobler a, Davison a, Pybus O.G., Kapan D.D., Planas S., Linares M., Heckel D., McMillan W.O. 2004. Historical demography of Mullerian mimicry in the neotropical Heliconius butterflies. Proc. Natl. Acad. Sci. U. S. A. 101:9704–9.

Fontaine M.C., Pease J.B., Steele A., Waterhouse R.M., Neafsey D.E., Sharakhov I. V., Jiang X., Hall A.B., Catteruccia F., Kakani E., Mitchell S.N., Wu Y.-C., Smith H.A., Love R.R., Lawniczak M.K., Slotman M.A., Emrich S.J., Hahn M.W., Besansky N.J. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science (80):science.1258524.

Fordyce J. a. 2010. Host shifts and evolutionary radiations of butterflies. Proc. Biol. Sci. 277:3735–43.

Forest F. 2009. Calibrating the Tree of Life: fossils, molecules and evolutionary timescales. Ann. Bot. 104:789–94.

Forister M.L., Novotny V., Panorska A.K., Baje L., Basset Y., Butterill P.T., Cizek L., Coley P.D., Dem F., Diniz I.R., Drozd P., Fox M., Glassmire A.E., Hazen R., Hrcek J., Jahner J.P., Kaman O., Kozubowski T.J., Kursar T.A., Lewis O.T., Lill J., Marquis R.J., Miller S.E., Morais H.C., Murakami M., Nickel H., Pardikes N.A., Ricklefs R.E., Singer M.S., Smilanich A.M., Stireman J.O., Villamarín-Cortez S., Vodka S., Volf M., Wagner D.L., Walla T., Weiblen G.D., Dyer L.A. 2015. The global distribution of diet breadth in insect herbivores. Proc. Natl. Acad. Sci. U. S. A. 112:442–7.

Fossilworks.org, (2015). *Fossilworks: Gateway to the Paleobiology Database*. [online] Available at: http://fossilworks.org [Accessed 1 Jun. 2015].

Fox L.R., Morrow P.A. 1981. Specialization: species property or local phenomenon? Science. 211:887–93.

Frantz A.C., Zachos F.E., Kirschning J., Cellina S., Bertouille S., Mamuris Z., Koutsogiannouli E.A., Burke T. 2013. Genetic evidence for introgression between domestic pigs and wild boars (Sus scrofa) in Belgium and Luxembourg: a comparative approach with multiple marker systems. Biol. J. Linn. Soc. 110:104–115.

Fuková I., Traut W., Vítková M., Nguyen P., Kubícková S., Marec F. 2007. Probing the W chromosome of the codling moth, Cydia pomonella, with sequences from microdissected sex chromatin. Chromosoma. 116:135–45.

Fulton T.L., Strobeck C. 2009. Multiple markers and multiple individuals refine true seal phylogeny and bring molecules and morphology back in line. Proc. R. Soc. B Biol. Sci. 277:1065–1070.

Gallant J.R., Imhoff V.E., Martin A., Savage W.K., Chamberlain N.L., Pote B.L., Peterson C., Smith G.E., Evans B., Reed R.D., Kronforst M.R., Mullen S.P. 2014. Ancient homology underlies adaptive mimetic diversity across butterflies. Nat. Commun. 5:4817.

Garrigan D., Kingan S.B., Geneva A.J., Andolfatto P., Clark A.G., Thornton K.R., Presgraves D.C. 2012. Genome sequencing reveals complex speciation in the Drosophila simulans clade. Genome Res. 22:1499–511.

Gatesy J., Baker R.H. 2005. Hidden likelihood support in genomic data: can forty-five wrongs

make a right? Syst. Biol. 54:483–92.

Gatesy J., Springer M.S. 2013. Concatenation versus coalescence versus “concatalescence”. Proc. Natl. Acad. Sci. U. S. A. 110:E1179.

Gatesy J., Springer M.S. 2014. Phylogenetic Analysis at Deep Timescales: Unreliable Gene Trees, Bypassed Hidden Support, and the Coalescence/Concatalescence Conundrum. Mol. Phylogenet. Evol. 80:231–266.

Gavrilets S., Losos J. 2009. Adaptive radiation: contrasting theory with data. Science. 323:732-737.

Gayral P., Melo-Ferreira J., Glémin S., Bierne N., Carneiro M., Nabholz B., Lourenco J.M., Alves P.C., Ballenghien M., Faivre N., Belkhir K., Cahais V., Loire E., Bernard A., Galtier N. 2013. Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. PLoS Genet. 9:e1003457.

Gbif.org, (2015). *Free and Open Access to Biodiversity Data | GBIF.org*. [online] Available at: http://www.gbif.org [Accessed 1 Jun. 2015].

Genome 10K Community of Scientists. 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J. Hered. 100:659–74.

Gerard D., Gibbs H.L., Kubatko L. 2011. Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. BMC Evol. Biol. 11:291.

Gilbert L.E. 1991. Biodiversity of a Central American Heliconius community: pattern, process, and problems. In: Price P., Lewinsohn T., Fernandes T., Benson W., editors. Plant-Animal Interactions: Evolutionary Ecology in Tropical and Temperate Regions. New York: John Wiley & Sons. p. 403–427.

Gilbert L.E. 2003. Adaptive novelty through introgression in Heliconius wings patterns: evidence for shared genetic “tool box” from synthetic hybrid zones and a theory of dievrsification. In: Boggs C.L., Watt W.B., Ehrlich P.R., editors. Butterflies: Ecology and evolution taking flight. Chicago: University of Chicago Press.

Glor R.E. 2010. Phylogenetic Insights on Adaptive Radiation. Annu. Rev. Ecol. Evol. Syst. 41:251–270.

Gompert Z., Lucas L.K., Buerkle C.A., Forister M.L., Fordyce J.A., Nice C.C. 2014. Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. Mol. Ecol. 23:4555–73.

Gompert Z., Lucas L.K., Nice C.C., Buerkle C.A. 2013. Genome divergence and the genetic architecture of barriers to gene flow between Lycaeides idas and L. Melissa. Evolution. 67:2498–514.

Gordon A. 2009. FASTX-Toolkit.

Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H.,

Zhai W., Fritz M.H.-Y., Hansen N.F., Durand E.Y., Malaspinas A.-S., Jensen J.D., Marques-Bonet T., Alkan C., Prüfer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-Petri A., Butthof A., Höber B., Höffner B., Siegemund M., Weihmann A., Nusbaum C., Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic D., Kucan Z., Gusic I., Doronichev V.B., Golovanova L. V, Lalueza-Fox C., de la Rasilla M., Fortea J., Rosas A., Schmitz R.W., Johnson P.L.F., Eichler E.E., Falush D., Birney E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D., Pääbo S. 2010. A draft sequence of the Neandertal genome. Science. 328:710–22.

Gregor H.-J. 1978. Die Miozaenan Frucht-und Samen-Floren die Oberpfaelzer Braunkohle. I. Funde aus den sandigen zwischenmitteln. Palaeontogr. Abteilung B. 167:8–103.

Gregory-Wodzicki K.M. 2000. Uplift history of the Central and Northern Andes: A review. Geol. Soc. Am. Bull. 112:1091–1105.

Greminger M.P., Stölting K.N., Nater A., Goossens B., Arora N., Bruggmann R., Patrignani A., Nussberger B., Sharma R., Kraus R.H.S., Ambu L.N., Singleton I., Chikhi L., van Schaik C.P., Krützen M. 2014. Generation of SNP datasets for orangutan population genomics using improved reduced-representation sequencing and direct comparisons of SNP calling algorithms. BMC Genomics. 15:16.

Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–21.

Hall J.P.W.J.P.W., Harvey D.J. 2002. The phylogeography of Amazonia revisited: New evidence from riodinid butterflies. Evolution (N. Y). 56:1489–1497.

Hallström B.M., Janke A. 2010. Mammalian evolution may not be strictly bifurcating. Mol. Biol. Evol. 27:2804–16.

Hand B.K., Hether T.D., Kovach R.P., Muhlfeld C.C., Amish S.J. 2015. Genomics and introgression: Discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. Curr. Zool. 61:146–154.

Harmon L.J., Weir J.T., Brock C.D., Glor R.E., Challenger W. 2008. GEIGER: investigating evolutionary radiations. Bioinformatics. 24:129–31.

Heath T.A., Hedtke S.M., Hillis D.M. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J. Syst. Evol. 46:239–257.

Heath T.A., Huelsenbeck J.P., Stadler T. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. Proc. Natl. Acad. Sci. 111:E2957–2966.

Hedrick, P.W. 2013. Adaptive introgression in animals. Mol. Eco. 22(18):4606-18.

Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–80.

Heliconius Genome Consortium. 2012. Islands of divergence underlie adaptive radiation in a butterfly genome. Nature. 487:94–98.

Hennig F., Meyer A. 2014. The evolutionary genomics of cichlid fishes: Explosive speciation and adaptation in the postgenomic era. *Annual Reviews of Genomics and Human Genetics.*15: 471-441.

Hermansen J.S., Haas F., Trier C.N., Bailey R.I., Nederbragt A.J., Marzal A., Saetre G.-P. 2014. Hybrid speciation through sorting of parental incompatibilities in Italian sparrows. Mol. Ecol. 23:5831–42.

Herrig D.K., Modrick A.J., Brud E., Llopart A. 2014. Introgression in the Drosophila subobscura--D. Madeirensis sister species: evidence of gene flow in nuclear genes despite mitochondrial differentiation. Evolution. 68:705–19.

Hewitson W.C. 1854. Illustrations of new species of exotic butterflies : selected chiefly from the collections of W. Wilson Saunders and William C. Hewitson. Vol. 1. London: John van Voorst.

Hey J. 2010. Isolation with migration models for more than two populations. Mol. Biol. Evol. 27:905–20.

Hill R.I., Gilbert L.E., Kronforst M.R. 2013. Cryptic genetic and wing pattern diversity in a mimetic Heliconius butterfly. Mol. Ecol. 22:2760–70.

Hillis D.M., Heath T.A., St. John K., John K. 2005. Analysis and Visualization of Tree Space. Syst. Biol. 54:471 – 482.

Hines H.M., Counterman B.A., Papa R., Albuquerque de Moura P., Cardoso M.Z., Linares M., Mallet J., Reed R.D., Jiggins C.D., Kronforst M.R., McMillan W.O., Albuquerque P., Moura D. 2011. Wing patterning gene redefines the mimetic history of Heliconius butterflies. Proc. Natl. Acad. Sci. U. S. A. 108:19666–71.

Hohenlohe P.A., Day M.D., Amish S.J., Miller M.R., Kamps-Hughes N., Boyer M.C., Muhlfeld C.C., Allendorf F.W., Johnson E.A., Luikart G. 2013. Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. Mol. Ecol. 22:3002–13.

Holland B.R., Huber K.T., Dress A., Moulton V. 2002. Delta plots: a tool for analyzing phylogenetic distance data. Mol. Biol. Evol. 19:2051–9.

Holzinger H., Holzinger R. 2000. Heliconius and related genera. Sciences Nat.

Hoorn C., Wesselingh F.P., ter Steege H., Bermudez M.A., Mora A., Sevink J., Sanmartín I., Sanchez-Meseguer A., Anderson C.L., Figueiredo J.P., Jaramillo C., Riff D., Negri F.R., Hooghiemstra H., Lundberg J., Stadler T., Särkinen T., Antonelli A. 2010. Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. Science. 330:927–31.

Huber B., Whibley A., Poul Y.L., Navarro N., Martin A., Baxter S., Shah A., Gilles B., Wirth T., McMillan W.O., Joron M. 2015. Conservatism and novelty in the genetic architecture of adaptation in Heliconius butterflies. Heredity (Edinb).

Huerta-Sánchez E., Jin X., Bianba Z., Peter B.M., Vinckenbosch N., Liang Y., Yi X., He M., Somel M., Ni P., Wang B., Ou X., Luosang J., Cuo Z.X.P., Li K., Gao G., Yin Y., Wang W., Zhang X., Xu X., Yang H., Li Y., Wang J., Wang J., Nielsen R. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 512:194–197.

Hull D. 1988. Science as a Process. University of Chicago Press. Chicago.

Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–67.

Huson D.H., Klöpper T., Lockhart P.J., Steel M.A. 2005. Reconstruction of Reticulate Networks from Gene Trees. Res. Comput. Mol. Biol. Lect. Notes Comput. Sci. Vol. 3500.:233–249.

Janz N., Nyblom K., Nylin S. 2001. Evolutionary dynamics of host-plant specialization: a case study of the tribe Nymphalini. Evolution. 55:783–96.

Janz N., Nylin S., Wahlberg N. 2006. Diversity begets diversity: host expansions and the diversification of plant-feeding insects. BMC Evol. Biol. 6:4.

Janz N., Wahlberg N. 2006. Diversity begets diversity : host expansions and the diversification of plant-feeding insects. 10:1–10.

Janzen D.H. 1980. When is it Coevolution? Evolution (N. Y). 34:611–612.

Janzen D.H. 1983. Erblichia odorata Seem. (Turneraceae) is a larval hostplant of *Eueides procula* vulgiformis (Nymphalidae: Heliconiini) in Santa Rosa National Park, Costa Rica. J. Lepid. Soc. 37.

Jaramillo C., Hoorn C., Silva S.A.F., Leite F., Herrera F., Quiroz L., Rodolfo D., Antonioli L. 2010. The origin of the modern Amazon rainforest: implications of the palynological and palaeobotanical record. In: Hoorn C., Wesselingh F.P., editors. Amazonia, Landscape and Species Evolution: A Look into the Past. Oxford: Blackwell. p. 317–334.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldon T., Capella-Gutierrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C. V., Lovell P. V., Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M. V., Alfaro-Nunez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M.,

Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jonsson K.A., Johnson W., Koepfli K.-P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alstrom P., Edwards S. V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science (80-.). 346:1320–1331.

Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–31.

Jiggins C.D., Linares M., Naisbit R.E., Salazar C., Yang Z.H., Mallet J. 2001a. Sex-linked hybrid sterility in a butterfly. Evolution (N. Y). 55:1631.

Jiggins C.D., Mcmillan W., King P., Mallet J. 1997a. The maintenance of species differences across a Heliconius hybrid zone. Heredity (Edinb). 79:495–505.

Jiggins C.D., McMillan W.O., Mallet J. 1997b. Host plant adaptation has not played a role in the recent speciation of Heliconius himera and Heliconius erato. Ecol. Entomol. 22:361–365.

Jiggins C.D., Naisbit R.E., Coe R.L., Mallet J. 2001b. Reproductive isolation caused by colour pattern mimicry. Nature. 411:302–5.

Jiggins C.D., Salazar C., Linares M., Mavarez J. 2008. Hybrid trait speciation and Heliconius butterflies. Philos. Trans. R. Soc. B Biol. Sci. 363:3047–3054.

Jiggins C.D. 2008a. Ecological Speciation in Mimetic Butterflies. Bioscience. 58:541.

Jiggins C.D. 2008b. Ecological Speciation in Mimetic Butterflies. Bioscience. 58:541.

Jombart T., Ahmed I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 27:3070–1.

Jombart T., Devillard S., Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11:94.

Jones J.C., Fan S., Franchini P., Schartl M., Meyer A. 2013a. The evolutionary history of Xiphophorus fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. Mol. Ecol. 22:2986–3001.

Jones R.T., Le Poul Y., Whibley A.C., Mérot C., ffrench-Constant R.H., Joron M. 2013b. Wing shape variation associated with mimicry in butterflies. Evolution. 67:2323–34.

Jorge L.R., Cordeiro-Estrela P., Klaczko L.B., Moreira G.R.P., Freitas A.V.L. 2011. Host-plant dependent wing phenotypic variation in the neotropical butterfly Heliconius erato. Biol. J. Linn. Soc. 102:765–774.

Joron M., Frezal L., Jones R.T., Chamberlain N.L., Lee S.F., Haag C.R., Whibley A., Becuwe

M., Baxter S.W., Ferguson L., Wilkinson P. a., Salazar C., Davidson C., Clark R., Quail M. a., Beasley H., Glithero R., Lloyd C., Sims S., Jones M.C., Rogers J., Jiggins C.D., ffrench-Constant R.H. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature. 477:203–6.

Kaellersjoe M., Farris J.J.S., Chase M.M.W., Bremer B., Fay M.F.M., Humphries C.J.C., Petersen G., Seberg O., Bremer K. 1997. Simultaneous parsimony jackknife analysis of 2538 rbcl dna sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. Plant Syst. Evol. 213:259–287.

Kang J.H., Schartl M., Walter R.B., Meyer A. 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 Evol. Biol. 13:25.

Kapan D.D., Flanagan N.S., Tobler A., Papa R., Reed R.D., Gonzalez J.A., Restrepo M.R., Martinez L., Maldonado K., Ritschoff C., Heckel D.G., McMillan W.O. 2006. Localization of Müllerian mimicry genes on a dense linkage map of Heliconius erato. Genetics. 173:735–57.

Katoh K. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.

Kawahara A.Y., Breinholt J.W. 2014. Phylogenomics provides strong evidence for relationships of butterflies and moths. Proc. Biol. Sci. 281:20140970.

Kayserili M.A., Gerrard D.T., Tomancak P., Kalinka A.T. 2012. An excess of gene expression divergence on the X chromosome in Drosophila embryos: implications for the faster-X hypothesis. PLoS Genet. 8:e1003200.

Keightley P.D., Pinharanda A., Ness R.W., Simpson F., Dasmahapatra K.K., Mallet J., Davey J.W., Jiggins C.D. 2014. Estimation of the spontaneous mutation rate in Heliconius melpomene. Mol. Biol. Evol.:msu302–.

Keller I., Bensasson D., Nichols R.A. 2007. Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. PLoS Genet. 3:e22.

Kelly E. 2009. A note on the name Hermathena and its lepidopterna namesakes. Hermathena. A Trinity Coll. Dublin Rev. 186.

Kembel S.W., Cowan P.D., Helmus M.R., Cornwell W.K., Morlon H., Ackerly D.D., Blomberg S.P., Webb C.O. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics. 26:1463–4.

Khadem M., Camacho R., Nóbrega C. 2011. Studies of the species barrier between Drosophila subobscura and D. madeirensis V: the importance of sex-linked inversion in preserving species identity. J. Evol. Biol. 24:1263–73.

Killip E.P. 1938. The American speciesof Passifloraceae. Publ. Field. Mus. Nat. Hist., Bot.

Ser. 19:1–613.

Kloepper T.H., Huson D.H. 2008. Drawing explicit phylogenetic networks and their integration into SplitsTree. BMC Evol. Biol. 8:22.

Knapp S., Mallet J. 1998. A New Species of Passiflora (Passifloraceae) from Ecuador with Notes on the Natural History of Its Herbivore, Heliconius (Lepidoptera: Nymphalidae: Heliconiiti). Novon. 8:162–166.

Knowles L., Kubatko L. 2010. Estimating species trees. Estimating species trees: practical and theoretical aspects. Hoboken: John Wiley & Sons. p. 1–15.

Kozak K.M., Wahlberg N., Neild A., Dasmahapatra K.K., Mallet J., Jiggins C.D. 2015. Multilocus Species Trees Show the Recent Adaptive Radiation of the Mimetic Heliconius Butterflies. Syst. Biol.:syv007–.

Kraus R.H.S., Kerstens H.H.D., van Hooft P., Megens H.-J., Elmberg J., Tsvey A., Sartakov D., Soloviev S.A., Crooijmans R.P.M.A., Groenen M.A.M., Ydenberg R.C., Prins H.H.T. 2012. Widespread horizontal genomic exchange does not erode species barriers among sympatric ducks. BMC Evol. Biol. 12:45.

Kronforst M.R., Salazar C., Linares M., Gilbert L.E. 2007. No genomic mosaicism in a putative hybrid butterfly species. Proc. Biol. Sci. 274:1255–64.

Kronforst M.R., Young L.G., Blume L.M., Gilbert L.E. 2006. Multilocus analyses of admixture and introgression among hybridizing Heliconius butterflies. Evolution. 60:1254–68.

Kronforst M.R.R., Hansen M.E.B.E.B., Crawford N.G.G., Gallant J.R.R., Zhang W., Kulathinal R.J.J., Kapan D.D.D., Mullen S.P.P. 2013. Hybridization Reveals the Evolving Genomic Architecture of Speciation. Cell Rep. 5:666–77.

Krosnick S.E., Ford A.J., Freudenstein J. V. 2009. Taxonomic revision of Passiflora subgenus Tetrapathea including the monotypic genera Hollrungia and Tetrapathea (Passifloraceae), and a new species of Passiflora. Syst. Bot. 34:375–385.

Krosnick S.E., Freudenstein J. V. 2005. Monophyly and Floral Character Homology of Old World Passiflora (Subgenus Decaloba: Supersection Disemma). Syst. Bot. 30:139–152.

Krosnick S.E., Porter-Utley K.E., MacDougal J.M., Jørgensen P.M., McDade L.A. 2013. New Insights into the Evolution of <I>Passiflora</I> subgenus <I>Decaloba</I> (Passifloraceae): Phylogenetic Relationships and Morphological Synapomorphies. Syst. Bot. 38:692–713.

Krosnick S.E. 2006. Phylogenetic relationships and patterns of morphological evolution in the Old World Species of Passiflora (Subgenus Decaloba: Supersection Disemma and Subgenus Tetrapathaea). PhD thesis.

Kubatko L., Meng C. 2010. Accomodating hybridisation in a multilocus phylogenetic

framework. Estimating species trees: practical and theoretical aspects. Hoboken: John Wiley & Sons. p. 99–114.

Kulathinal R.J., Stevison L.S., Noor M.A.F. 2009. The genomics of speciation in Drosophila: diversity, divergence, and introgression estimated using low-coverage genome sequencing. PLoS Genet. 5:e1000550.

Kumar S., Filipski A.J., Battistuzzi F.U., Kosakovsky Pond S.L., Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457–72.

Kunte K., Shea C., Aardema M.L., Scriber J.M., Juenger T.E., Gilbert L.E., Kronforst M.R. 2011. Sex chromosome mosaicism and hybrid speciation among tiger swallowtail butterflies. PLoS Genet. 7:e1002274.

Kunte K., Zhang W., Tenger-Trolander A., Palmer D.H., Martin A., Reed R.D., Mullen S.P., Kronforst M.R. 2014. doublesex is a mimicry supergene. Nature. 507:229–32.

Lamas G., Callaghan C., Casagrande M., Mielke O., Pyrcz T., Robbins R., Viloria. A. 2004. Hesperioidea -- Papilionoidea. In: Heppner J., editor. Atlas of Neotropical Lepidoptera. Checklist: part 4A. Gainesville, Florida: Association for Tropical Lepidoptera/Scientific Publishers.

Lamichhaney S., Berglund J., Almén M.S., Maqbool K., Grabherr M., Martinez-Barrio A., Promerová M., Rubin C.-J., Wang C., Zamani N., Grant B.R., Grant P.R., Webster M.T., Andersson L. 2015. Evolution of Darwin’s finches and their beaks revealed by genome sequencing. Nature. 518:371–375.

Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29:1695–701.

Lanier H.C., Huang H., Knowles L.L. 2014. How low can you go? The effects of mutation rate on the accuracy of species-tree estimation. Mol. Phylogenet. Evol. 70:112–9.

Larget B.R., Kotha S.K., Dewey C.N., Ané C. 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics. 26:2910–1.

Lavoie C.A., Platt R.N., Novick P.A., Counterman B.A., Ray D.A. 2013. Transposable element evolution in Heliconius suggests genome diversity within Lepidoptera. Mob. DNA. 4:21.

Leaché A.D., Harris R.B., Rannala B., Yang Z. 2013. The Influence of Gene Flow on Species Tree Estimation: A Simulation Study. Syst. Biol.

Leaché A.D., Rannala B. 2010. The Accuracy of Species Tree Estimation under Simulation: A Comparison of Methods. Syst. Biol.:1–12.

Leaché A.D., Wagner P., Linkem C.W., Böhme W., Papenfuss T.J., Chong R.A., Lavin B.R., Bauer A.M., Nielsen S. V, Greenbaum E., Rödel M.-O., Schmitz A., LeBreton M., Ineich

I., Chirio L., Ofori-Boateng C., Eniang E.A., Baha El Din S., Lemmon A.R., Burbrink F.T. 2014. A hybrid phylogenetic-phylogenomic approach for species tree estimation in African Agama lizards with applications to biogeography, character evolution, and diversification. Mol. Phylogenet. Evol. 79:215–230.

Lee C.S., McCool B.A., Moore J.L., Hillis D.M., Gilbert L.E. 1992. Phylogenetic study of heliconiine butterflies based on morphology and restriction analysis of ribosomal RNA genes. Zool. J. Linn. Soc. 106:17–31.

Lee Y., Marsden C.D., Nieman C., Lanzaro G.C. 2014. A new multiplex SNP genotyping assay for detecting hybridization and introgression between the M and S molecular forms of Anopheles gambiae. Mol. Ecol. Resour. 14:297–305.

Lee Y., Marsden C.D., Norris L.C., Collier T.C., Main B.J., Fofana A., Cornel A.J., Lanzaro G.C. 2013. Spatiotemporal dynamics of gene flow and hybrid fitness between the M and S forms of the malaria mosquito, Anopheles gambiae. Proc. Natl. Acad. Sci. U. S. A. 110:19854–9.

Legendre P., Desdevises Y., Bazin E. 2002. A statistical test for host-parasite coevolution. Syst. Biol. 51:217–34.

Leigh J.W., Susko E., Baumgartner M., Roger A.J. 2008. Testing congruence in phylogenomic analysis. Syst. Biol. 57:104–15.

Lemmon A.R., Brown J.M., Stanger-Hall K., Lemmon E.M. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58:130–45.

Lemmon A.R., Lemmon E.M. 2012. High-throughput identification of informative nuclear loci for shallow-scale phylogenetics and phylogeography. Syst. Biol. 61:745–61.

Lewis A.R., Marchant D.R., Ashworth A.C., Hemming S.R., Machlus M.L. 2007. Major middle Miocene global climate change: Evidence from East Antarctica and the Transantarctic Mountains. Geol. Soc. Am. Bull. 119:1449–1461.

Li H., Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25:1754–60.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–9.

Lima T.G. 2014. Higher levels of sex chromosome heteromorphism are associated with markedly stronger reproductive isolation. Nat. Commun. 5:4743.

Linnen C.R., Farrell B.D. 2008. Comparison of methods for species-tree inference in the sawfly genus Neodiprion (Hymenoptera: Diprionidae). Syst. Biol. 57:876–90.

Liu K., Linder C.R., Warnow T. 2011. RAxML and FastTree: comparing two methods for

large-scale maximum likelihood phylogeny estimation. PLoS One. 6:e27731.

Liu L., Yu L., Edwards S. V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10:302.

Lunt D.H., Kumar S., Koutsovoulos G., Blaxter M.L. 2014. The complex hybrid origins of the root knot nematodes revealed through comparative genomics. PeerJ. 2:e356.

Lunter G., Goodson M. 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 21:936–9.

Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.

Maddison W.P., Maddison D.R. 2011. Mesquite: a modular system for evolutionary analysis.

Maddison W.P. 1997. Gene Trees in Species Trees. Syst. Biol. 46:523–536.

Mai D.H. 1967. Die florenzonen der florenwechsel und die vorstellungen über den klimaablauf im Jungteriär der Deutschen Demokratischen Republik. Abhandlungen des Zent. Geol. Institutes. 10:55–81.

Mallet J., Barton N.H. 1989. Strong Natural Selection in a Warning-Color Hybrid Zone. Evolution (N. Y). 43:421.

Mallet J., Beltrán M., Neukirchen W., Linares M. 2007. Natural hybridization in heliconiine butterflies: the species boundary as a continuum. BMC Evol. Biol. 7:28.

Mallet J., Gilbert L.E. 1995. Why are there so many mimicry rings? Correlations between habitat, behaviour and mimicry in Heliconius butterflies. Biol. J. Linn. Soc. 55:159–180.

Mallet J., McMillan W., Jiggins C. 1998. Mimicry and warning color at the boundary between races and species. In: Howard D., Berlocher S., editors. Endless Forms: Species and Speciation. New York: Oxford Univ. Press. p. 390–403.

Mallet J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20:229–37.

Martin A., Orgogozo V. 2013. The Loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. Evolution. 67:1235–50.

Martin A., Papa R., Nadeau N.J., Hill R.I., Counterman B.A., Halder G., Jiggins C.D., Kronforst M.R., Long A.D., McMillan W.O., Reed R.D. 2012. Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand. Proc. Natl. Acad. Sci. U. S. A. 109:12632–7.

Martin S.H., Dasmahapatra K.K., Nadeau N.J., Salazar C., Walters J.R., Simpson F., Blaxter M., Manica A., Mallet J., Jiggins C.D. 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 23:1817–28.

Martin S.H., Davey J.W., Jiggins C.D. 2015a. Evaluating the Use of ABBA-BABA Statistics to Locate Introgressed Loci. Mol. Biol. Evol.:msu269–.

Martin S.H., Eriksson A., Kozak K.M., Manica A., Jiggins C.D. 2015b. Speciation in Heliconius butterflies: Minimal contact followed by millions of generations of hybridisation. Submitted.

Massardo D., Fornel R., Kronforst M., Gonçalves G.L., Moreira G.R.P. 2014. Diversification of the silverspot butterflies (Nymphalidae) in the Neotropics inferred from multi-locus DNA sequences. Mol. Phylogenet. Evol.

Matasci N., Hung L.-H., Yan Z., Carpenter E.J., Wickett N.J., Mirarab S., Nguyen N., Warnow T., Ayyampalayam S., Barker M., Burleigh J., Gitzendanner M.A., Wafula E., Der J.P., dePamphilis C.W., Roure B., Philippe H., Ruhfel B.R., Miles N.W., Graham S.W., Mathews S., Surek B., Melkonian M., Soltis D.E., Soltis P.S., Rothfels C., Pokorny L., Shaw J.A., DeGironimo L., Stevenson D.W., Villarreal J., Chen T., Kutchan T.M., Rolf M., Baucom R.S., Deyholos M.K., Samudrala R., Tian Z., Wu X., Sun X., Zhang Y., Wang J., Leebens-Mack J., Wong G.K.-S. 2014. Data access for the 1,000 Plants (1KP) project. Gigascience. 3:17.

Mavárez J., Salazar C. a, Bermingham E., Salcedo C., Jiggins C.D., Linares M. 2006. Speciation by hybridization in Heliconius butterflies. Nature. 441:868–71.

McCormack J.E., Hird S.M., Zellmer A.J., Carstens B.C., Brumfield R.T. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol. 66:526–38.

McDade L.A. 2000. Hybridization and phylogenetics: Special insights from morphology. In: Wiens J.J., editor. Morphological Data in Phylogenetic Analysis: Recent Progress and Unresolved Problems. Washington, D.C.: Smithsonian Institution Press.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M.A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–303.

McKenna D.D., Sequeira A.S., Marvaldi A.E., Farrell B.D. 2009. Temporal lags and overlap in the diversification of weevils and flowering plants. Proc. Natl. Acad. Sci. U. S. A. 106:7083–8.

Meier-Kolthoff J.P., Auch A.F., Huson D.H., Göker M. 2007. COPYCAT: cophylogenetic analysis tool. Bioinformatics. 23:898–900.

Melo M.C., Salazar C., Jiggins C.D., Linares M. 2009. Assortative mating preferences among hybrids offers a route to hybrid speciation. Evolution. 63:1660–5.

Mendez F.L., Watkins J.C., Hammer M.F. 2012. Global genetic variation at OAS1 provides evidence of archaic admixture in Melanesian populations. Mol. Biol. Evol. 29:1513–20.

Merian M.S. 1705. Insects of Surinam (Metamorphosis Insectorum Surinamensium). Amsterdam: Johann Georg Volckamer.

Mérot C., Mavárez J., Evin A., Dasmahapatra K.K., Mallet J., Lamas G., Joron M. 2013. Genetic differentiation without mimicry shift in a pair of hybridizing Heliconius species (Lepidoptera: Nymphalidae). Biol. J. Linn. Soc. 109:830–847.

Merrill R.M., Naisbit R.E., Mallet J., Jiggins C.D. 2013. Ecological and genetic factors influencing the transition between host-use strategies in sympatric Heliconius butterflies. J. Evol. Biol. 26:1959–67.

Merrill R.M., Van Schooten B., Scott J.A., Jiggins C.D. 2011. Pervasive genetic associations between traits causing reproductive isolation in Heliconius butterflies. Proc. Biol. Sci. 278:511–8.

Merrill R.M., Wallbank R.W.R., Bull V., Salazar P.C.A., Mallet J., Stevens M., Jiggins C.D. 2012. Disruptive ecological selection on a mating cue. Proc. Biol. Sci. 279:4907–13.

Merrill R.M., et al. 2015. The diversification of *Heliconius* butterflies: What have we learned in 150 years? J. Evo. Bio. DOI: 10.1111/jeb.12672

Meyer M., Kircher M., Gansauge M.-T., Li H., Racimo F., Mallick S., Schraiber J.G., Jay F., Prüfer K., de Filippo C., Sudmant P.H., Alkan C., Fu Q., Do R., Rohland N., Tandon A., Siebauer M., Green R.E., Bryc K., Briggs A.W., Stenzel U., Dabney J., Shendure J., Kitzman J., Hammer M.F., Shunkov M. V, Derevianko A.P., Patterson N., Andrés A.M., Eichler E.E., Slatkin M., Reich D., Kelso J., Pääbo S. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science. 338:222–6.

Mirarab S., Bayzid M.S., Boussau B., Warnow T. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science (80-.). 346:1250463–1250463.

Misof B., Liu S., Meusemann K., Peters R.S., Donath A., Mayer C., Frandsen P.B., Ware J., Flouri T., Beutel R.G., Niehuis O., Petersen M., Izquierdo-Carrasco F., Wappler T., Rust J., Aberer A.J., Aspock U., Aspock H., Bartel D., Blanke A., Berger S., Bohm A., Buckley T.R., Calcott B., Chen J., Friedrich F., Fukui M., Fujita M., Greve C., Grobe P., Gu S., Huang Y., Jermiin L.S., Kawahara A.Y., Krogmann L., Kubiak M., Lanfear R., Letsch H., Li Y., Li Z., Li J., Lu H., Machida R., Mashimo Y., Kapli P., McKenna D.D., Meng G., Nakagaki Y., Navarrete-Heredia J.L., Ott M., Ou Y., Pass G., Podsiadlowski L., Pohl H., von Reumont B.M., Schutte K., Sekiya K., Shimizu S., Slipinski A., Stamatakis A., Song W., Su X., Szucsich N.U., Tan M., Tan X., Tang M., Tang J., Timelthaler G., Tomizuka S., Trautwein M., Tong X., Uchifune T., Walzl M.G., Wiegmann B.M., Wilbrandt J., Wipfler B., Wong T.K.F., Wu Q., Wu G., Xie Y., Yang S., Yang Q., Yeates D.K., Yoshizawa K., Zhang Q., Zhang R., Zhang W., Zhang Y., Zhao J., Zhou C., Zhou L., Ziesmann T., Zou S., Xu X., Yang H., Wang J., Kjer K.M., Zhou X. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science (80-.). 346:763–767.

Moen D., Morlon H. 2014. From dinosaurs to modern bird diversity: extending the time scale of adaptive radiation. PLoS Biol. 12:e1001854.

Monzón J., Kays R., Dykhuizen D.E. 2014. Assessment of coyote-wolf-dog admixture using ancestry-informative diagnostic SNPs. Mol. Ecol. 23:182–97.

Moreira G.R.P., Mielke C.G.C. 2010. A new species of Neruda Turner, 1976 from northeast Brazil (Lepidoptera: Nymphalidae, Heiconiinae, Heliconiini). Nachrichten des Entomol. Vereins Apollo. 31:85–91.

Müller C.J., Beheregaray L.B. 2010. Palaeo island-affinities revisited--biogeography and systematics of the Indo-Pacific genus Cethosia Fabricius (Lepidoptera: Nymphalidae). Mol. Phylogenet. Evol. 57:314–26.

Muschner V.C., Zamberlan P.M., Bonatto S.L., Freitas L.B. 2012. Phylogeny , biogeography and divergence times in Passiflora (Passifloraceae). 4:1036–1043.

Muschner V.C. 2003. A first molecular phylogenetic analysis of Passiflora (Passifloraceae). Am. J. Bot. 90:1229–1238.

Nadeau N., Pardo-Diaz C., Whibley A., Supple M.A., Wallbank R., Wu G.C., Maroja L., Ferguson L., Hines H., Salazar C., ffrench-Constant R., Joron M., McMillan W.O., Jiggins C. 2015. The origins of a novel butterfly wing patterning gene from within a family of conserved cell cycle regulators. bioRxiv.:016006.

Nadeau N.J., Martin S.H., Kozak K.M., Salazar C., Dasmahapatra K.K., Davey J.W., Baxter S.W., Blaxter M.L., Mallet J., Jiggins C.D. 2013. Genome-wide patterns of divergence and gene flow across a butterfly radiation. Mol. Ecol. 22:814–26.

Nadeau N.J., Whibley A., Jones R.T., Davey J.W., Dasmahapatra K.K., Baxter S.W., Quail M.A., Joron M., ffrench-Constant R.H., Blaxter M.L., Mallet J., Jiggins C.D. 2012. Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 367:343–53.

Naisbit R.E., Jiggins C.D., Linares M., Salazar C., Mallet J. 2002. Hybrid sterility, Haldane's rule and speciation in Heliconius cydno and H. melpomene. Genetics. 161:1517–26.

Nakhleh L. 2013. Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol. Evol. 28:719–28.

Neafsey D.E., Waterhouse R.M., Abai M.R., Aganezov S.S., Alekseyev M.A., Allen J.E., Amon J., Arca B., Arensburger P., Artemov G., Assour L.A., Basseri H., Berlin A., Birren B.W., Blandin S.A., Brockman A.I., Burkot T.R., Burt A., Chan C.S., Chauve C., Chiu J.C., Christensen M., Costantini C., Davidson V.L.M., Deligianni E., Dottorini T., Dritsou V., Gabriel S.B., Guelbeogo W.M., Hall A.B., Han M. V., Hlaing T., Hughes D.S.T., Jenkins A.M., Jiang X., Jungreis I., Kakani E.G., Kamali M., Kemppainen P., Kennedy R.C., Kirmitzoglou I.K., Koekemoer L.L., Laban N., Langridge N., Lawniczak M.K.N., Lirakis M., Lobo N.F., Lowy E., MacCallum R.M., Mao C., Maslen G., Mbogo C., McCarthy J., Michel K., Mitchell S.N., Moore W., Murphy K.A., Naumenko A.N., Nolan T., Novoa E.M., O'Loughlin S., Oringanje C., Oshaghi M.A., Pakpour N., Papathanos P.A., Peery A.N., Povelones M., Prakash A., Price D.P., Rajaraman A.,

Reimer L.J., Rinker D.C., Rokas A., Russell T.L., Sagnon N., Sharakhova M. V., Shea T., Simao F.A., Simard F., Slotman M.A., Somboon P., Stegniy V., Struchiner C.J., Thomas G.W.C., Tojo M., Topalis P., Tubio J.M.C., Unger M.F., Vontas J., Walton C., Wilding C.S., Willis J.H., Wu Y.-C., Yan G., Zdobnov E.M., Zhou X., Catteruccia F., Christophides G.K., Collins F.H., Cornman R.S., Crisanti A., Donnelly M.J., Emrich S.J., Fontaine M.C., Gelbart W., Hahn M.W., Hansen I.A., Howell P.I., Kafatos F.C., Kellis M., Lawson D., Louis C., Luckhart S., Muskavitch M.A.T., Ribeiro J.M., Riehle M.A., Sharakhov I. V., Tu Z., Zwiebel L.J., Besansky N.J. 2014. Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. Science (80). 347:1258522.

Ness R.W., Morgan A.D., Colegrave N., Keightley P.D. 2012. Estimate of the spontaneous mutation rate in Chlamydomonas reinhardtii. Genetics. 192:1447–54.

Newberry J.S. 1895. The flora of the Amboy clays. Washington: USGS.

Nice C.C., Gompert Z., Fordyce J.A., Forister M.L., Lucas L.K., Buerkle C.A. 2013. Hybrid speciation and independent evolution in lineages of alpine butterflies. Evolution. 67:1055–68.

Nijhout F. 1991. The Development And Evolution Of Butterfly Wing Patterns. Washington & London: Smithsonian Institution Press.

Nishikawa H., Iijima T., Kajitani R., Yamaguchi J., Ando T., Suzuki Y., Sugano S., Fujiyama A., Kosugi S., Hirakawa H., Tabata S., Ozaki K., Morimoto H., Ihara K., Obara M., Hori H., Itoh T., Fujiwara H. 2015. A genetic mechanism for female-limited Batesian mimicry in Papilio butterfly. *Nature Genetics* **47**, 405–409 .

Norris L.C., Main B.J., Lee Y., Collier T.C., Fofana A., Cornel A.J., Lanzaro G.C. 2015. Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets. Proc. Natl. Acad. Sci. 112:201418892.

Nosil P. 2012. Ecological Speciation. Oxford, UK: Oxford University Press.

Nylander J.A.A. 2004. MrModelTest.

Ocampo J., d'Eeckenbrugge G.C., Jarvis A. 2010. Distribution of the Genus Passiflora L. Diversity in Colombia and Its Potential as an Indicator for Biodiversity Management in the Coffee Growing Zone. Diversity. 2:1158–1180.

Occhipinti A. 2013. Plant coevolution: evidences and new challenges. J. Plant Interact. 8:188–196.

Papa R., Kapan D.D., Counterman B.A., Maldonado K., Lindstrom D.P., Reed R.D., Nijhout H.F., Hrbek T., McMillan W.O. 2013. Multi-allelic major effect genes interact with minor effect QTLs to control adaptive color pattern variation in Heliconius erato. PLoS One. 8:e57033.

Papa R., Morrison C.M., Walters J.R., Counterman B.A., Chen R., Halder G., Ferguson L., Chamberlain N., Ffrench-Constant R., Kapan D.D., Jiggins C.D., Reed R.D., Mcmillan W.O. 2008. Highly conserved gene order and numerous novel repetitive elements in genomic regions linked to wing pattern variation in Heliconius butterflies. BMC Genomics. 9:345.

Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 20:289–90.

Parchman T.L., Gompert Z., Braun M.J., Brumfield R.T., McDonald D.B., Uy J.A.C., Zhang G., Jarvis E.D., Schlinger B.A., Buerkle C.A. 2013. The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. Mol. Ecol. 22:3304–17.

Pardo-Diaz C., Salazar C., Baxter S.W., Merot C., Figueiredo-Ready W., Joron M., McMillan W.O., Jiggins C.D. 2012. Adaptive Introgression across Species Boundaries in Heliconius Butterflies. PLoS Genet. 8:e1002752.

Parham J.F., Donoghue P.C.J., Bell C.J., Calway T.D., Head J.J., Holroyd P.A., Inoue J.G., Irmis R.B., Joyce W.G., Ksepka D.T., Patané J.S.L., Smith N.D., Tarver J.E., van Tuinen M., Yang Z., Angielczyk K.D., Greenwood J.M., Hipsley C.A., Jacobs L., Makovicky P.J., Müller J., Smith K.T., Theodor J.M., Warnock R.C.M., Benton M.J. 2012. Best practices for justifying fossil calibrations. Syst. Biol. 61:346–59.

Pease J.B., Hahn M.W. 2013. More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. Evolution. 67:2376–84.

Pemberton R. 1989. Insects attacking Passiflora mollissima and other Passiflora species; field survey in the Andes. Proc. Hawaiian Entomol. Soc. 29:71–84.

Penney H.D., Hassall C., Skevington J.H., Abbott K.R., Sherratt T.N. 2012. A comparative analysis of the evolution of imperfect mimicry. Nature. 483:461–4.

Penz C. 1999. Higher level phylogeny for the passion-vine butterflies (Nymphalidae, Heliconiinae) based on early stage and adult morphology. Zool. J. Linn. Soc. 127:277–344.

Penz C.M., Peggie D. 2003. Phylogenetic relationships among Heliconiinae genera based on morphology (Lepidoptera: Nymphalidae). Syst. Entomol. 28:451–479.

Pfennig D., Editor G. 2012. Mimicry: Ecology, evolution, and development. Curr. Zool. 58:604–606.

Pickrell J.K., Pritchard J.K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8:e1002967.

Plummer M., Best N., Cowles K., Vines K. 2006. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News. 6.

Poelstra J.W., Vijay N., Bossu C.M., Lantz H., Ryll B., Müller I., Baglione V., Unneberg P., Wikelski M., Grabherr M.G., Wolf J.B.W. 2014. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science. 344:1410–4.

Pohl N., Sison-Mangus M.P., Yee E.N., Liswi S.W., Briscoe A.D. 2009. Impact of duplicate gene copies on phylogenetic analysis and divergence time estimates in butterflies. BMC Evol. Biol. 9:99.

Porter-Utley K. 2014. A revision of PassifloraL.subgenusDecaloba(DC.) Rchb.supersectionCieca (Medik.) J. M. MacDougal & Feuillet (Passifloraceae). PhytoKeys.:1–224.

Posada D., Crandall K.A. 1998. MODELTEST: testing the model of DNA substitution. Evolution (N. Y). 14:817–818.

Posada D. 2008. jModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25:1253–6.

Price M.N., Dehal P.S., Arkin A.P. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 5:e9490.

Prowell D.P. 1998. Sex linkage and speciation in Lepidoptera. In: Howard D.J., Berlocher S., editors. Endless Forms: Species and Speciation. Oxford: Oxford University Press.

Pruitt K.D., Tatusova T., Maglott D.R. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 33:D501–4.

Puigbò P., Lobkovsky A.E., Kristensen D.M., Wolf Y.I., Koonin E. V. 2014. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC Biol. 12:66.

Pulquério M.J.F., Nichols R.A. 2007. Dates from the molecular clock: how wrong can we be? Trends Ecol. Evol. 22:180–4.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A.R., Bender D., Maller J., Sklar P., de Bakker P.I.W., Daly M.J., Sham P.C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–75.

Purcell S. 2009. PLINK.

Pybus O.G., Harvey P.H. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc. Biol. Sci. 267:2267–72.

Pyron R.A., Hendry C.R., Chou V.M., Lemmon E.M., Lemmon A.R., Burbrink F.T. 2014. Effectiveness of phylogenomic data and coalescent species-tree methods for resolving difficult nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia). Mol. Phylogenet. Evol. 81:221–31.

Quah S., Hui J.H.L., Holland P.W.H. 2015. A burst of miRNA innovation in the early evolution of butterflies and moths. Mol. Biol. Evol.:msv004–.

Quek S.-P., Counterman B. a, Albuquerque de Moura P., Cardoso M.Z., Marshall C.R., McMillan W.O., Kronforst M.R. 2010. Dissecting comimetic radiations in Heliconius reveals divergent histories of convergent butterflies. Proc. Natl. Acad. Sci. U. S. A. 107:7365–70.

R Development Core Team. 2008. R: A language and environment for statistical computing.

Rabosky D.L., Donnellan S.C., Grundler M., Lovette I.J. 2014a. Analysis and visualization of complex macroevolutionary dynamics: an example from Australian scincid lizards. Syst. Biol. 63:610–27.

Rabosky D.L., Grundler M., Anderson C., Title P., Shi J.J., Brown J.W., Huang H., Larson J.G. 2014b. BAMMtools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. Methods Ecol. Evol. 5:n/a–n/a.

Rabosky D.L. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. PLoS One. 9:e89543.

Rambaut A., Suchard M., Xie W., Drummond A. 2014. Tracer v. 1.6.

Rambaut A. 2009. FigTree. Tree figure drawing tool.

Ranwez V., Delsuc F., Ranwez S., Belkhir K., Tilak M.-K., Douzery E.J. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. BMC Evol. Biol. 7:241.

Rasky K. 1960. Pflanzenreste aus dem Obereozän Ungrans. Senckenbergiana Lethaea. 41:423–449.

Rauscher M. 1988. Is coevolution dead? Ecology. 69:898–901.

Reed R.D., Mcmillan W.O., Nagy L.M., B P.R.S. 2009. Gene expression underlying adaptive variation in Heliconius wing patterns : non-modular regulation of overlapping cinnabar and vermilion prepatterns Gene expression underlying adaptive variation in Heliconius wing patterns : non-modular regulation of over. October.

Reed R.D., Nagy L.M. 2005. Evolutionary redeployment of a biosynthetic module: expression of eye pigment genes vermilion, cinnabar, and white in butterfly wing development. Evol. Dev. 7:301–11.

Regier J.C., Shultz J.W., Zwick A., Hussey A., Ball B., Wetzer R., Martin J.W., Cunningham C.W. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature. 463:1079–83.

Reich D., Green R.E., Kircher M., Krause J., Patterson N., Durand E.Y., Viola B., Briggs A.W., Stenzel U., Johnson P.L.F., Maricic T., Good J.M., Marques-Bonet T., Alkan C., Fu Q., Mallick S., Li H., Meyer M., Eichler E.E., Stoneking M., Richards M., Talamo S., Shunkov M. V, Derevianko A.P., Hublin J.-J., Kelso J., Slatkin M., Pääbo S. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature.

468:1053–60.

Reid N.M., Hird S.M., Brown J.M., Pelletier T.A., McVay J.D., Satler J.D., Carstens B.C. 2014. Poor fit to the multispecies coalescent is widely detectable in empirical data. Syst. Biol. 63:322–33.

Renaut S. 2011. Contemporary hybrid speciation in sculpins (Cottus spp.). Mol. Ecol. 20:1320–1.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Robinson G.E., Hackett K.J., Purcell-Miramontes M., Brown S.J., Evans J.D., Goldsmith M.R., Lawson D., Okamuro J., Robertson H.M., Schneider D.J. 2011. Creating a buzz about insect genomes. Science. 331:1386.

Rodrigues D., Moreira G.R.P. 2002. Geographical variation in larval host-plant use by Heliconius erato (Lepidoptera: Nymphalidae) and consequences for adult life history. Brazilian J. Biol. 62:321–332.

Ronquist F., Huelsenbeck J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Rosser N., Phillimore A.B., Huertas B., Willmott K.R., Mallet J. 2012. Testing historical explanations for gradients in species richness in heliconiine butterflies of tropical America. Biol. J. Linn. Soc. 105:479–497.

Rosser N., Kozak K., Phillimore A., Mallet J. 2015. Extensive range overlap between *Heliconius* sister species: evidence for sympatric speciation in butterflies? BMC Evolutionary Biology. In press.

Rota-Stabelli O., Campbell L., Brinkmann H., Edgecombe G.D., Longhorn S.J., Peterson K.J., Pisani D., Philippe H., Telford M.J. 2010. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc. R. Soc. B Biol. Sci. 278(1703):298-306.

Roure B., Baurain D., Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30:197–214.

Roux C., Tsagkogeorga G., Bierne N., Galtier N. 2013. Crossing the species barrier: genomic hotspots of introgression between two highly divergent Ciona intestinalis species. Mol. Biol. Evol. 30:1574–87.

Rull V. 2011. Neotropical biodiversity: timing and potential drivers. Trends Ecol. Evol. 26:508–513.

Sahara K., Yoshido A., Traut W. 2012. Sex chromosome evolution in moths and butterflies. Chromosome Res. 20:83–94.

Salazar C., Baxter S.W., Pardo-Diaz C., Wu G., Surridge A., Linares M., Bermingham E.,

Jiggins C.D. 2010. Genetic evidence for hybrid trait speciation in heliconius butterflies. PLoS Genet. 6:e1000930.

Salazar C., Jiggins C.D., Taylor J.E., Kronforst M.R., Linares M. 2008. Gene flow and the genealogical history of Heliconius heurippa. BMC Evol. Biol. 8:132.

Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 497:327–31.

Salichos L., Stamatakis A., Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31:1261–71.

Salzberg S.L., Phillippy A.M., Zimin A., Puiu D., Magoc T., Koren S., Treangen T.J., Schatz M.C., Delcher A.L., Roberts M., Marçais G., Pop M., Yorke J. a. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. 22:557–67.

Sankararaman S., Mallick S., Dannemann M., Prüfer K., Kelso J., Pääbo S., Patterson N., Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 507:354–7.

Sauquet H., Ho S.Y.W., Gandolfo M.A., Jordan G.J., Wilf P., Cantrill D.J., Bayly M.J., Bromham L., Brown G.K., Carpenter R.J., Lee D.M., Murphy D.J., Sniderman J.M.K., Udovicic F. 2012. Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of Nothofagus (Fagales). Syst. Biol. 61:289–313.

Savage W.K., Mullen S.P. 2009. A single origin of Batesian mimicry among hybridizing populations of admiral butterflies (Limenitis arthemis) rejects an evolutionary reversion to the ancestral phenotype. Proc. Biol. Sci. 276:2557–65.

Savolainen V., Chase M.W., Hoot S.B., Morton C.M., Soltis D.E., Bayer C., Fay M.F., De Bruijn A.Y., Sullivan S., Qiu Y.-L. 2000. Phylogenetics of Flowering Plants Based on Combined Analysis of Plastid atpB and rbcL Gene Sequences. Syst. Biol. 49:306–362.

Scally A., Dutheil J.Y., Hillier L.W., Jordan G.E., Goodhead I., Herrero J., Hobolth A., Lappalainen T., Mailund T., Marques-Bonet T., McCarthy S., Montgomery S.H., Schwalie P.C., Tang Y.A., Ward M.C., Xue Y., Yngvadottir B., Alkan C., Andersen L.N., Ayub Q., Ball E. V, Beal K., Bradley B.J., Chen Y., Clee C.M., Fitzgerald S., Graves T.A., Gu Y., Heath P., Heger A., Karakoc E., Kolb-Kokocinski A., Laird G.K., Lunter G., Meader S., Mort M., Mullikin J.C., Munch K., O'Connor T.D., Phillips A.D., Prado-Martinez J., Rogers A.S., Sajjadian S., Schmidt D., Shaw K., Simpson J.T., Stenson P.D., Turner D.J., Vigilant L., Vilella A.J., Whitener W., Zhu B., Cooper D.N., de Jong P., Dermitzakis E.T., Eichler E.E., Flicek P., Goldman N., Mundy N.I., Ning Z., Odom D.T., Ponting C.P., Quail M.A., Ryder O.A., Searle S.M., Warren W.C., Wilson R.K., Schierup M.H., Rogers J., Tyler-Smith C., Durbin R. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature. 483:169–75.

Schluter D. 2000. The Ecology of Adaptive Radiation. Oxford: Oxford University Press.

Schumer M., Cui R., Boussau B., Walter R., Rosenthal G., Andolfatto P. 2013. An evaluation of the hybrid speciation hypothesis for Xiphophorus clemenciae based on whole genome sequences. Evolution. 67:1155–68.

Schumer M., Rosenthal G.G., Andolfatto P. 2014. How common is homoploid hybrid speciation? Evolution. 68:1553–60.

Schwarz G. 1978. Estimating the Dimension of a Model. Ann. Stat. 6:461–464.

Seehausen O., Butlin R.K., Keller I., Wagner C.E., Boughman J.W., Hohenlohe P.A., Peichel C.L., Saetre G.-P., Bank C., Brännström A., Brelsford A., Clarkson C.S., Eroukhmanoff F., Feder J.L., Fischer M.C., Foote A.D., Franchini P., Jiggins C.D., Jones F.C., Lindholm A.K., Lucek K., Maan M.E., Marques D.A., Martin S.H., Matthews B., Meier J.I., Möst M., Nachman M.W., Nonaka E., Rennison D.J., Schwarzer J., Watson E.T., Westram A.M., Widmer A. 2014. Genomics and the origin of species. Nat. Rev. Genet. 15:176–92.

Ševcik J., Kvacek Z., Mai D. 2007. A new mastixioid florula from tektite-bearing deposits in South Bohemia, Czech Republic (Middle Miocene, Vrábče Member). Bull. Geosci. 82:429 – 436.

Shaw T.I., Ruan Z., Glenn T.C., Liu L. 2013. STRAW: Species TRee Analysis Web server. Nucleic Acids Res. 41:W238–41.

Sheppard P.M., Turner J.R.G., Brown K.S., Benson W.W., Singer M.C. 1985. Genetics and the Evolution of Muellerian Mimicry in Heliconius Butterflies. Philos. Trans. R. Soc. B Biol. Sci. 308:433–610.

Sherratt T.N. 2008. The evolution of Müllerian mimicry. Naturwissenschaften. 95:681–95.

Shimizu A., Dohzono I., Nakaji M., Roff D.A., Miller D.G., Osato S., Yajima T., Niitsu S., Utsugi N., Sugawara T., Yoshimura J. 2014. Fine-tuned bee-flower coevolutionary state hidden within multiple pollination interactions. Sci. Rep. 4:3988.

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

Silva A.K., Gonçalves G.L., Moreira G.R.P. 2014. Larval feeding choices in heliconians: induced preferences are not constrained by performance and host plant phylogeny. Anim. Behav. 89:155–162.

Silvério A., de Araujo Mariath J.E. 2013. Comparative structure of the pollen in species of Passiflora: insights from the pollen wall and cytoplasm contents. Plant Syst. Evol. 300:347–358.

Simpson J.T., Wong K., Jackman S.D., Schein J.E., Jones S.J.M., Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117–23.

Smith B.T., Harvey M.G., Faircloth B.C., Glenn T.C., Brumfield R.T. 2013. Target Capture and Massively Parallel Sequencing of Ultraconserved Elements for Comparative Studies

at Shallow Evolutionary Time Scales. Syst. Biol. doi: 10.1093/sysbio/syt061

Solomon S.E., Bacci M., Martins J., Vinha G.G., Mueller U.G. 2008. Paleodistributions and comparative molecular phylogeography of leafcutter ants (Atta spp.) provide new insight into the origins of Amazonian diversity. PLoS One. 3:e2738.

Soltis D.E., Smith S.A., Cellinese N., Wurdack K.J., Tank D.C., Brockington S.F., Refulio-Rodriguez N.F., Walker J.B., Moore M.J., Carlsward B.S., Bell C.D., Latvis M., Crawley S., Black C., Diouf D., Xi Z., Rushworth C.A., Gitzendanner M.A., Sytsma K.J., Qiu Y.-L., Hilu K.W., Davis C.C., Sanderson M.J., Beaman R.S., Olmstead R.G., Judd W.S., Donoghue M.J., Soltis P.S. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Am. J. Bot. 98:704–30.

Song S., Liu L., Edwards S. V, Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. U. S. A. 109:14942–7.

Spencer K. 1988. Chemical mediation of coevolution in Passiflora-Heliconius interaction. In: Spencer K.C., editor. Chemical mediation of coevolution. American Institute of Biological Sciences.

Stadler T. 2009. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. J. Theor. Biol. 261:58–66.

Stadler T. 2013. Recovering speciation and extinction dynamics based on phylogenies. J. Evol. Biol. 26:1203–19.

Stamatakis A., Auch A.F., Meier-Kolthoff J., Göker M. 2007. AxPcoords & parallel AxParafit: statistical co-phylogenetic analyses on thousands of taxa. BMC Bioinformatics. 8:405.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–90.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–3.

Staubach F., Lorenc A., Messer P.W., Tang K., Petrov D.A., Tautz D. 2012. Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (Mus musculus). PLoS Genet. 8:e1002891.

Steel M. 2005. Should phylogenetic models be trying to “fit an elephant”? Trends Genet. 21:307–9.

Stemshorn K.C., Reed F.A., Nolte A.W., Tautz D. 2011. Rapid formation of distinct hybrid lineages after secondary contact of two fish species (Cottus sp.). Mol. Ecol. 20:1475–1491.

Strimmer K., Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. Proc. Biol. Sci. 269:137–42.

Suetsugu Y., Futahashi R., Kanamori H., Kadono-Okuda K., Sasanuma S., Narukawa J., Ajimura M., Jouraku A., Namiki N., Shimomura M., Sezutsu H., Osanai-Futahashi M., Suzuki M.G., Daimon T., Shinoda T., Taniai K., Asaoka K., Niwa R., Kawaoka S., Katsuma S., Tamura T., Noda H., Kasahara M., Sugano S., Suzuki Y., Fujiwara H., Kataoka H., Arunkumar K.P., Tomar A., Nagaraju J., Goldsmith M.R., Feng Q., Xia Q., Yamamoto K., Shimada T., Mita K. 2013. Large scale full-length cDNA sequencing reveals a unique genomic landscape in a lepidopteran model insect, Bombyx mori. G3 (Bethesda). 3:1481–92.

Supple M.A., Hines H.M., Dasmahapatra K.K., Lewis J.J., Nielsen D.M., Lavoie C., Ray D.A., Salazar C., McMillan W.O., Counterman B.A. 2013. Genomic architecture of adaptive color pattern divergence and convergence in Heliconius butterflies. Genome Res. 23:1248–57.

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

Takahata N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics. 122:957–66.

Tamura K., Battistuzzi F.U., Billing-Ross P., Murillo O., Filipski A., Kumar S. 2012. Estimating divergence times in large molecular phylogenies. Proc. Natl. Acad. Sci. U. S. A. 109:19333–8.

Tamura K., Stecher G., Peterson D., Filipski A., Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. 30:2725–9.

Teeter K.C., Thibodeau L.M., Gompert Z., Buerkle C.A., Nachman M.W., Tucker P.K. 2010. The variable genomic architecture of isolation between hybridizing species of house mice. Evolution. 64:472–85.

Teo S.M., Pawitan Y., Ku C.S., Chia K.S., Salim A. 2012. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics. 28:2711–8.

Than C., Ruths D., Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics. 9:322.

Thiergart T., Landan G., Martin W.F. 2014. Concatenated alignments and the case of the disappearing tree. BMC Evol. Biol. 14:2624.

Thomas J.A., Trueman J.W.H., Rambaut A., Welch J.J. 2013. Relaxed phylogenetics and the palaeoptera problem: resolving deep ancestral splits in the insect phylogeny. Syst. Biol. 62:285–97.

Thompson J. 1999. What we know and do not know about coevolution: insect herbivores and plants as a test case. In: Olff H., Brown V.K., Drent R.H., editors. Herbivores: between plants and predators. Oxford, UK: Blackwell Science Ltd. p. 7–30.

Thomson R.C., Shaffer H.B. 2010. Sparse supermatrices for phylogenetic inference:

taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. Syst. Biol. 59:42–58.

Timmermans M.J.T.N., Lees D.C., Simonsen T.J. 2014. Towards a mitogenomic phylogeny of Lepidoptera. Mol. Phylogenet. Evol. 79:169–78.

Tokuoka T. 2012. Molecular phylogenetic analysis of Passifloraceae sensu lato (Malpighiales) based on plastid and nuclear DNA sequences. J. Plant Res. 125:489–97.

Tolweb.org, (2015). *Tree of Life Web Project*. [online] Available at: http://tolweb.org [Accessed 7 Nov. 2015].

Traut W., Niimi T., Ikeo K., Sahara K. 2006. Phylogeny of the sex-determining gene Sex-lethal in insects. Genome. 49:254–62.

Traut W., Vogel H., Glöckner G., Hartmann E., Heckel D.G. 2013. High-throughput sequencing of a single chromosome: a moth W chromosome. Chromosom. Res. 21:491–505.

Trier C.N., Hermansen J.S., Sætre G.-P., Bailey R.I. 2014. Evidence for mito-nuclear and sex-linked reproductive barriers between the hybrid Italian sparrow and its parent species. PLoS Genet. 10:e1004075.

Turchetto-Zolet A.C., Pinheiro F., Salgueiro F., Palma-Silva C. 2013. Phylogeographical patterns shed light on evolutionary process in South America. Mol. Ecol. 22:1193–213.

Turner J. 1965. Evolution of complex polymorphism and mimicry in distasteful South American butterflies. Proc. XII Int. Cong. Entomol. London.:267.

Turner J.R., Johnson M.S., Eanes W.F. 1979. Contrasted modes of evolution in the same genome: allozymes and adaptive change in Heliconius. Proc. Natl. Acad. Sci. U. S. A. 76:1924–8.

Twyford A., Ennos R. 2012. Next-generation hybridization and introgression. Heredity (Edinb). 108:179–89.

Ulmer T., MacDougal J.M. 2004. Passiflora. Passion flowers of the world. Cambridge, UK.: Timber Press.

Vamosi S.M. 2005. On the role of enemies in divergence and diversification of prey: a review and synthesis. Can. J. Zool. 83:894–910.

van't Hof A.E., Edmonds N., Dalíková M., Marec F., Saccheri I.J. 2011. Industrial melanism in British peppered moths has a singular and recent mutational origin. Science. 332:958–60.

Velasco J., Steel M. 2014. Axiomatic opportunities and obstacles for inferring a species tree from gene trees. Syst. Biol. 63:772–778.

Van Velzen R., Wahlberg N., Sosef M.S.M., Bakker F.T. 2013. Effects of changing climate on species diversification in tropical forest butterflies of the genus Cymothoe (Lepidoptera:

Nymphalidae). Biol. J. Linn. Soc. 108:546–564.

Venables W.N., Ripley B.D. 2003. Modern applied statistics with S. Springer.

vonHoldt B.M., Pollinger J.P., Earl D.A., Knowles J.C., Boyko A.R., Parker H., Geffen E., Pilot M., Jedrzejewski W., Jedrzejewska B., Sidorovich V., Greco C., Randi E., Musiani M., Kays R., Bustamante C.D., Ostrander E.A., Novembre J., Wayne R.K. 2011. A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. Genome Res. 21:1294–305.

Wagner C.E., Keller I., Wittwer S., Selz O.M., Mwaiko S., Greuter L., Sivasundar A., Seehausen O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22:787–98.

Wahlberg N., Leneveu J., Kodandaramaiah U., Peña C., Nylin S., Freitas A.V.L., Brower A.V.Z. 2009. Nymphalid butterflies diversify following near demise at the Cretaceous/Tertiary boundary. Proc. Biol. Sci. 276:4295–302.

Wallbank R., Baxter S., Pardo-Diaz C., Hanly J., Martin S., Mallet J., Dasmahapatra K., Joron M., Nadeau N., McMillan W., Jiggins C. 2015. The origins of an evolutionary novelty through modular regulation of an input-output gene. *In review.*

Weetman D., Wilding C.S., Steen K., Pinto J., Donnelly M.J. 2012. Gene flow-dependent genomic divergence between Anopheles gambiae M and S forms. Mol. Biol. Evol. 29:279–91.

Weitemier K., Straub S.C.K., Cronn R.C., Fishbein M., Schmickl R., McDonnell A., Liston A. 2014. Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant Phylogenomics. Appl. Plant Sci. 2:1400042.

Welch J.J., Jiggins C.D. 2014. Standing and flowing: the complex origins of adaptive variation. Mol. Ecol. 23:3935–7.

White M.A., Ané C., Dewey C.N., Larget B.R., Payseur B.A. 2009. Fine-scale phylogenetic discordance across the house mouse genome. PLoS Genet. 5:e1000729.

Wiens J.J., Morrill M.C. 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst. Biol. 60:719–31.

Wiens J.J. 2003. Missing Data, Incomplete Taxa, and Phylogenetic Accuracy. Syst. Biol. 52:528–538.

Wright J.J. 2011. Conservative coevolution of Müllerian mimicry in a group of rift lake catfish. Evolution. 65:395–407.

Wu L.-W., Lin L.-H., Lees D.C., Hsu Y.-F. 2014. Mitogenomic sequences effectively recover relationships within brush-footed butterflies (Lepidoptera: Nymphalidae). BMC Genomics. 15:468.

Wurdack K.J., Davis C.C. 2009. Malpighiales phylogenetics: Gaining ground on one of the most recalcitrant clades in the angiosperm tree of life. Am. J. Bot. 96:1551–70.

Xi Z., Ruhfel B.R., Schaefer H., Amorim A.M., Sugumaran M., Wurdack K.J., Endress P.K., Matthews M.L., Stevens P.F., Mathews S., Davis C.C. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proc. Natl. Acad. Sci. U. S. A. 109:17519–24.

Xia Q., Zhou Z., Lu C., Cheng D., Dai F., Li B., Zhao P., Zha X., Cheng T., Chai C., Pan G., Xu J.J., Liu C., Lin Y., Qian J., Hou Y., Wu Z., Li G.G.G., Pan M., Li C.C., Shen Y., Lan X., Yuan L., Li T., Xu H., Yang G., Wan Y., Zhu Y., Yu M., Shen W., Wu D., Xiang Z., Yu J., Wang J.J.J., Li R., Shi J., Li H., Su J., Wang X., Zhang Z.Z., Wu Q., Li J.J., Zhang Q., Wei N., Sun H., Dong L., Liu D., Zhao S., Zhao X., Meng Q., Lan F., Huang X., Li Y., Fang L., Li D., Sun Y., Yang Z., Huang Y., Xi Y., Qi Q., He D., Huang H., Zhang X., Wang Z., Li W., Cao Y., Yu Y., Yu H., Ye J.J., Chen H., Zhou Y., Liu B., Ji H., Li S.S., Ni P., Zhang J., Zhang Y., Zheng H., Mao B., Wang W., Ye C., Wong G.K.-S., Yang H. 2004. A draft sequence for the genome of the domesticated silkworm (Bombyx mori). Science. 306:1937–40.

Xia X., Xie Z. 2001. DAMBE: software package for data analysis in molecular biology and evolution. J. Hered. 92:371–3.

Yockteng R., Nadot S. 2004. Infrageneric phylogenies : a comparison of chloroplast-expressed glutamine synthetase , cytosol-expressed glutamine synthetase and cpDNA maturase K in Passiflora. 31:397–402.

You M., Yue Z., He W., Yang X., Yang G., Xie M., Zhan D., Baxter S.W., Vasseur L., Gurr G.M., Douglas C.J., Bai J., Wang P., Cui K., Huang S., Li X., Zhou Q., Wu Z., Chen Q., Liu C., Wang B., Li X., Xu X., Lu C., Hu M., Davey J.W., Smith S.M., Chen M., Xia X., Tang W., Ke F., Zheng D., Hu Y., Song F., You Y., Ma X., Peng L., Zheng Y., Liang Y., Chen Y., Yu L., Zhang Y., Liu Y., Li G., Fang L., Li J., Zhou X., Luo Y., Gou C., Wang J., Wang J., Yang H., Wang J. 2013. A heterozygous moth genome provides insights into herbivory and detoxification. Nat. Genet. 45:220–5.

Yu Y., Dong J., Liu K.J., Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proc. Natl. Acad. Sci. 111(46):16448-16453.

Yu Y., Ristic N., Nakhleh L. 2013. Fast algorithms and heuristics for phylogenomics under ILS and hybridization. BMC Bioinformatics. 14 Suppl 1:S6.

Yu Y., Than C., Degnan J.H., Nakhleh L. 2011. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Syst. Biol. 60:138–49.

Zaman L., Meyer J.R., Devangam S., Bryson D.M., Lenski R.E., Ofria C. 2014. Coevolution Drives the Emergence of Complex Traits and Promotes Evolvability. PLoS Biol. 12:e1002023.

Zhan S., Merlin C., Boore J.L., Reppert S.M. 2011. The monarch butterfly genome yields

insights into long-distance migration. Cell. 147:1171–85.

Zhan S., Zhang W., Niitepõld K., Hsu J., Haeger J.F., Zalucki M.P., Altizer S., de Roode J.C., Reppert S.M., Kronforst M.R. 2014. The genetics of monarch butterfly migration and warning colouration. Nature. 514:317–321.

Zhang J., Kapli P., Pavlidis P., Stamatakis A. 2013a. A general species delimitation method with applications to phylogenetic placements. Bioinformatics. 29:2869–76.

Zhang W., Kunte K., Kronforst M.R. 2013b. Genome-wide characterization of adaptation and speciation in tiger swallowtail butterflies using de novo transcriptome assemblies. Genome Biol. Evol. 5:1233–45.