Published in final edited form as:
Science. 2019 November 01; 366(6465): 594–599. doi:10.1126/science.aaw2090.

Genomic architecture and introgression shape a butterfly radiation

Nathaniel B. Edelman[1,*], Paul B. Frandsen[2,3], Michael Miyagi[1], Bernardo Clavijo[4], John Davey[5], Rebecca Dikow[3], Gonzalo García-Accinelli[4], Steven M. Van Belleghem[6], Nick Patterson[7,8], Daniel E. Neafsey[8,9], Richard Challis[10], Sujai Kumar[11], Gilson R. P. Moreira[12], Camilo Salazar[13], Mathieu Chouteau[14], Brian A. Counterman[15], Riccardo Papa[6,16], Mark Blaxter[10], Robert D. Reed[17], Kanchon K. Dasmahapatra[5], Marcus Kronforst[18], Mathieu Joron[19], Chris D. Jiggins[20], W. Owen McMillan[21], Federica Di Palma[4], Andrew J. Blumberg[22], John Wakeley[1], David Jaffe[8,23], James Mallet[1,*]

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, 02138, USA

[2]Department of Plant and Wildlife Sciences, Brigham Young University, Provo, UT, 84602, USA

[3]Data Science Lab, Office of the Chief Information Officer, Smithsonian Institution, Washington, D.C., 20560, USA

[4]Earlham Institute, Norwich Research Park, NR4 7UZ, UK

[5]Department of Biology, University of York, YO10 5DD, UK

[6]Department of Biology, University of Puerto Rico, Río Piedras Campus, San Juan, Puerto Rico

[7]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, 02138, USA

[8]Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA

[9]Harvard TH Chan School of Public Health, Boston, MA, 02115, USA

[10]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, CB10 1SA

[11]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, EH9 3JT, UK

[12]Departamento de Zoologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil

[*]Correspondence to: nedelman@g.harvard.edu; jmallet@oeb.harvard.edu.

Author contributions: NBE, NP, DEN, RC, SK, MB, RDR, KKD, MK, MJ, CDJ, WOM, DJ, and JM conceived the project; NBE, PBF, MM, BC, DJ, and JM designed experiments; JD, GM, CS, MC, BAC, RDR, KKD, MJ, CDJ, WOM, and JM supplied specimens and extracted DNA; DJ, BC, GGA, and FDP assembled genomes, PBF inferred species trees; NBE performed reference gap filling, phylogenetic network, local gene tree, recombination rate, and genome structure analyses; JD and SVB generated linkage maps, MM, JW, and AB developed the branch length test; NBE and JM wrote and edited the manuscript with input from all authors.

Competing interests: Authors declare no competing interests.

Data and materials availability: Reads generated are available in the SRA, BioProject PRJNA531399. Genome assemblies are available on LepBase (http://lepbase.org). The full multi-species, whole genome alignment, as well as sub-alignments for windows and their respective inferred phylogenies are available on Dryad, DOI 10.5061/dryad.b7bj832. Also on Dryad are configuration files for progressiveCactus, the inferred LD recombination map for *H.* erato, loci used for phyloNet, and repeatMasker output for all assemblies. All code used to analyze data is available on Zenodo repositories, DOIs 10.5281/zenodo.3286366 and 10.5281/zenodo.2653677.

[13]Biology Program, Faculty of Natural Sciences and Mathematics, Universidad del Rosario, Carrera 24, No. 63C-69, Bogotá, D.C. 111221, Colombia

[14]Laboratoire Ecologie, Evolution, Interactions des Systémes Amazoniens (LEEISA), USR 3456, Université De Guyane, CNRS Guyane, 275 Route de Montabo, 97334 Cayenne, French Guiana.

[15]Department of Biological Sciences, Mississippi State University, Starkville, Mississippi 39762, USA

[16]Molecular Sciences and Research Center, University of Puerto Rico, San Juan, Puerto Rico

[17]Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY 14853, USA

[18]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA

[19]CEFE, CNRS, Université de Montpellier, Université Paul Valéry Montpellier 3, EPHE, IRD, Montpellier, France

[20]Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK

[21]Smithsonian Tropical Research Institute, Apartado 0843-03092, Panamá, Panama

[22]Department of Mathematics, University of Texas, Austin, TX, 78712, USA

[23]10x Genomics, Pleasanton, California 94566, USA

Abstract

We use twenty *de novo* genome assemblies to probe the speciation history and architecture of gene flow in rapidly radiating *Heliconius* butterflies. Our tests to distinguish incomplete lineage sorting from introgression indicate that gene flow has obscured several ancient phylogenetic relationships in this group over large swathes of the genome. Introgressed loci are underrepresented in low recombination and gene-rich regions, consistent with the purging of foreign alleles more tightly linked to incompatibility loci. We identify a hitherto unknown inversion that traps a color pattern switch locus. We infer that this inversion was transferred between lineages via introgression and is convergent with a similar rearrangement in another part of the genus. These multiple *de novo* genome sequences enable improved understanding of the importance of introgression and selective processes in adaptive radiation.

One Sentence Summary

Introgression has been a major contributor of genealogical discordance throughout *Heliconius* evolution, varying across the genome with local recombination rate, gene density, and genome architecture.

Adaptive radiations play a fundamental role in generating biodiversity. Initiated by key innovations and ecological opportunity, radiation is fueled by niche competition that promotes rapid diversification of species (7). Reticulate evolution may enhance radiation by introducing genetic variation, enabling rapidly emerging populations to take advantage of novel ecological opportunities (2, 3). Diverging from its sister genus *Eueides* ~12 My ago, *Heliconius* radiated in a burst of speciation in the last ~5 My (4). Introgression is well known in *Heliconius*, with widespread reticulate evolution across the genus (5), though this has been disputed (6). Nonetheless, how introgression varies across the genome is known only in one pair of sister lineages (7, 8). Here, we use multiple *de novo* whole genome assemblies to improve the resolution of introgression, incomplete lineage sorting (ILS), and genome architecture in deeper branches of the *Heliconius* phylogeny.

Phylogenetic analysis

We generated 20 *de novo* genome assemblies for species in both major *Heliconius* sub-clades and three additional genera of Heliconiini. Here we align the sixteen highest quality Heliconiini assemblies to two *Heliconius* reference genomes and seven other Lepidoptera genomes, resulting in an alignment of 25 taxa (9). *De novo* assembly provides superior sequence information for low complexity regions, allows for discovery of structural rearrangements, and improves alignment of evolutionarily distant clades (10). Other studies in *Heliconius* have shown a high level of phylogenetic discordance, arguably a result of rampant introgression (4, 5). We attempted to reconstruct a bifurcating species tree by estimating relationships using protein-coding genes, conserved coding regions, and conserved non-coding regions. We generated phylogenies with coalescent-based and concatenation approaches, using both the full Lepidoptera alignment and a restricted, Heliconiini-only sub-alignment. These topologies were largely congruent among analytical approaches, but weakly supported nodes were resolved inconsistently. These approaches therefore failed to resolve the phylogeny of *Heliconius* as a simple bifurcating tree (Fig. 1A, Fig. S20).

To determine whether hybridization was a cause of the species tree uncertainty, we calculated Patterson's *D*-statistics (11) for every triplet of the 13 *Heliconius* species, using a member of the sister genus, *Eueides tales,* as outgroup. In 201 of 286 triplets, we observed values significantly different from zero based on block-jackknifing, demonstrating strong evidence for introgression (Fig. S53). However, these tests alone yield little quantitative information about admixture. We therefore used phyloNet (12) to infer reticulate phylogenetic networks of these species on the basis of random samples of one hundred 10 kb windows across the alignment. For each sample, we co-estimated all 100 regional gene trees and the overall species network in parallel (12). To improve alignments, we analyzed the *melpomene*-silvaniform group with respect to the *H. melpomene* Hmel2.5 assembly (13) and the *erato-sara* group with respect to the *H. erato demophoon* v1 assembly (9, 14). Most species exhibited an admixture event at some point in their history using this method; we confirmed extensive reticulation among silvaniform species and discovered major gene flow events in the *erato-sara* clade. Based on these results, we propose the reticulate phylogenies in Fig. 1B–C.

Correlation of local ancestry with genome architecture

We next analyzed the distribution of tree topologies across the genome, again treating each major clade separately and using its respective reference genome. The *melpomene*-silvaniform group lacked topological consensus, unsurprisingly since introgression, especially of key mimicry loci, is well known from this clade (15). The most common tree topology was found in only 4.3% of windows, with an additional 14 topologies appearing in

1.0–3.4% of windows (Fig. S19–Fig. S21). By contrast, we here focus on the *erato-sara* group, where two topologies dominate (Fig. 2). One (Tree 2, Fig. 2B) matched our bifurcating consensus topology (Fig. 1A) and a recently published tree (4), while the other (Tree 1) differs in that it places *H. hecalesia* and *H. telesiphe* as sisters.

Regions with local topologies discordant from the species tree may have arisen through introgression or ILS. In order to make within-topology locus-by-locus inferences, we developed a statistical test to distinguish between ILS and introgression based on the distribution of internal branch lengths among windows for a given three-taxon subtree, conditional on its topology. We call this method Quantifying Introgression via Branch Lengths (QuIBL). In the absence of introgression, we expect internal branch lengths of triplet topologies discordant with the species tree (due to ILS) to be exponentially distributed. However, if introgression has occurred, their distribution should have that same exponential component, but also include an additional component with a non-zero mode corresponding to the time between the introgression event and the most recent common ancestor of all three species (9). Like other tree-based methods, QuIBL is potentially sensitive to the assumption that each tree is inferred from loci with limited internal recombination (Fig. S75). We therefore chose small (5 kb) windows to reduce the probability of intra-locus recombination breakpoints.

For every triplet in the *erato-sara* clade, we calculate the likelihood that the distribution of internal branch lengths is consistent with introgression or with ILS only. We formally distinguish between these two models using a BIC test with a strict cutoff of ΔBIC > 10. Consistent with our results from *D*-statistics, we find that 13 of 20 triplets have evidence for introgression (Table S13). For example, using QuIBL on the triplet *H. erato-H. hecalesia-H. telesiphe,* we infer that 76% of discordant loci, or 38% of all loci genome-wide, are introgressed. Averaging over all triplets, we infer that 71% (67% with BIC filtering) of loci with discordant gene trees have a history of introgression, or 20% (19% with BIC filtering) of all triplet loci, indicating a broad signal of introgression throughout the clade (Equation 7.7, Table S13; see (9) for additional discussion).

In hybrid populations, individuals have genomic regions that originate from different species and may be incompatible with the recipient genome or with their environment (16). Linked selection causes harmless or even beneficial introgressed loci to be removed along with these deleterious loci if they are tightly linked; this effect depends on the strength of selection and the local recombination rate (17, 18). We therefore expect introgressed loci to be enriched in regions where selection is likely to be weak, such as gene deserts, or in regions of high recombination, where harmless introgressed loci more readily recombine away from linked incompatibility loci.

In *Heliconius*, even distant species like *H. erato* and *H. melpomene* have the same number of broadly collinear chromosomes (13), facilitating direct comparisons among species. Furthermore, each chromosome in *Heliconius* has approximately one crossover per chromosome per meiosis in males (there is no crossing over in female *Heliconius*) (14, 19). Chromosomes vary in length, and chromosome size is inversely proportional to recombination rate per base pair (8, 13). We found a strong correlation between the fraction of windows in each chromosome that show a given topology and physical chromosome length (Fig. 3A). Such relationships exist for all 8 trees in Fig. 2B (9), but we focus here on the two most common trees: Tree 1 has a strongly negative correlation with chromosome size (r^2=0.883, *t*= 11.7, 18 *d.f., p*<0.0001) while Tree 2 (concordant with our inferred species tree) has a positive correlation (r^2=0.726, *t*=6.9, 18 *d.f., p*<0.0001). Results from QuIBL indicate that 94% of windows that recover a Tree 1 triplet topology are consistent with introgression (Fig. S70, Table S13). The Z (sex) chromosome 21, is strongly enriched for Tree 2, suggesting it may harbor more incompatibility loci than autosomes. Interspecific hybrid females in *Heliconius* are often sterile, conforming to Haldane's Rule, and sex chromosomes have been implicated as particularly important in generating incompatibilities (8, 20–24).

To test whether the pattern we observe among chromosomes is related to differences in recombination, we investigated the relationship between recombination rate and tree topology within chromosomes. Recombination rate declines at the ends of chromosomes (Fig. S85), and the species tree (Tree 2) is more abundant in those regions (Fig. 3B). In addition, when windows are grouped by local recombination rate calculated from population genetic data (9, 14), we observe a strong relationship with the recovered topology (Fig. 3C). Finally, we observe a minor enrichment of Tree 1 in regions of very low gene density, but this effect is weak (Fig. 3D) compared to that of recombination. Taken together, these results show that tighter linkage on longer chromosomes, and in lower recombination regions within chromosomes leads to removal of more introgressed variation in those regions. This very strong correlation is consistent with a highly polygenic architecture of incompatibilities between species.

Introgression of a convergent inversion

The topology block size distribution in the *erato* clade generally decayed exponentially (Fig. 2C), but two unusually long blocks contained minor topologies: one on chromosome 2 (Tree 3, composed of three sub-blocks) and the other on chromosome 15 (Tree 4). Our study of the ~3 Mb topology block on chromosome 2 confirms an earlier finding of an inversion in *H. erato* (13), and we show here that its rare topology is most likely explained by ILS including a long period of ancestral polymorphism (Fig. S95).

The topology block on chromosome 15 is of particular interest, as it spans *cortex*, a genetic hotspot of wing color pattern diversity in Lepidoptera (25, 26). We hypothesized that this block could be an inversion, as in *H. numata*, where the P_1 'supergene' inversion polymorphism around *cortex* controls color pattern switching among mimicry morphs (27). This block recovers *H. telesiphe* and *H. hecalesia* as a monophyletic subclade, which together are sister to the *sara* clade (Fig. 2B, Tree 4). We searched our *de novo* assemblies for contigs that mapped across topology transitions. Taking *H. melpomene* as the standard arrangement, we find clear inversion breakpoints in *H. telesiphe, H. hecalesia, H. sara,* and *H. demeter*. Conversely, *H. erato, H. himera,* and *E. tales* all contain contigs that map in their entirety across the breakpoints (Fig. 4A), implying that they have the ancestral *H. melpomene* arrangement.

This chromosome 15 inversion covers almost exactly the same region as the 400 kb P_1 inversion in *H. numata* (25, 27, 28). However, *de novo* contigs from our *H. numata* assembly show that the breakpoints of P_1 are close to but not identical to those of the inversion in the *erato* clade (Fig. 4A). Furthermore, in topologies for *H. numata, H. telesiphe, H. erato,* and *E. tales* across chromosome 15, not a single window recovered *H. numata* and *H. telesiphe* as a monophyletic subclade, as would be expected if the *erato* group inversion was homologous to P_1 in *H. numata*.

We used QuIBL with the triplet (*H. erato +H. telesiphe* + *H. sara*) to elucidate the evolutionary history of this inversion. A small internal branch would suggest ILS while a large internal branch would be more consistent with introgression (Fig. 4B). The average internal branch length in the inversion was much longer than the genome-wide average, corresponding to a 79% probability of introgression (Fig. 4C). If the inversion was polymorphic in the ancestral population for some time, we could also recover a similarly long internal branch (Fig. 4B, center). We distinguish between this longer-term polymorphic scenario and introgression by comparing the genetic distance (D_{xy}) between *H. telesiphe* and *H. sara,* represented by T_3 in Fig. 4B. Normalized D_{xy} (as in Fig. S95) within the inversion is ~25% less than in the rest of the genome. Given that this is a large genomic block, introgression is therefore the most parsimonious explanation for the evolutionary history of the inversion (Fig. 4D) (29).

Discussion

Species involved in rapid radiations are prone to hybridization due to frequent geographical overlap with closely related taxa. In both *melpomene* and *erato* clades of *Heliconius,* introgression has overwritten the original bifurcation history of several species across large swathes of the genome, a pattern also observed in *Anopheles* mosquitos (30). This observation is also consistent with genomic analysis of other rapid radiations characterized by widespread hybridization and introgression, including Darwin’s finches (2) and African cichlids (31). In other radiations, the role of introgression is less clear: in *Tamias* chipmunks, widespread introgression of mitochondrial DNA was identified, in contrast to an absence of evidence for nuclear gene flow (32). With few genomic comparisons available to date, it is perhaps too early to say whether introgression is a major feature of adaptive radiations in general, but evidence thus far suggests this to be the case.

Our results raise the question of why some genomic regions cross species boundaries while others do not. In the *erato* clade, we find a strong correlation between recombination rate and introgression probability. Similar associations with topology also exist between sister species in the *melpomene* clade (7). Associations between recombination and introgression in actively hybridizing populations of sword-tail fish (*Xiphophorus*) and monkey flowers (*Mimulus*) support the role of linked selection on a highly polygenic landscape of interspecific incompatibilities (18, 33, 34). Our results establish that this relationship persists and may indeed be strengthened with time since introgression. While hybridization is ongoing, many introgressed blocks are constantly reintroduced into the population. If linked to weakly deleterious alleles, introgressed loci will be finally purged by linked selection only long after introgression ceases.

Recombination rate alone cannot account for differential introgression, so we must delve into specific regions to elucidate their function and relevance to speciation. It is critical, therefore, to have tools that can confidently identify introgressed loci, and much effort has gone into developing such methods (11, 35). Our test using internal branch lengths in triplet gene trees is based in coalescent theory and takes advantage of the discriminatory power of a property of gene trees not explicitly accounted for by other methods. QuIBL allows us to assess probability of introgression for each locus in each species triplet (8). Here, we employ this method to identify the evolutionary origin of a convergent inversion that has undergone multiple independent introgression events, and to show that genomic regions with discordant topologies arose mostly through hybridization. Just as sex aids adaptation within species, occasional introgression and recombination among species can have major long-term effects on the genome, contributing variation that could fuel rapid adaptive divergence and radiation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Harvard FAS Research Computing team and the Smithsonian Institution for their support. Conversations with S. Martin were instrumental in developing our thinking around recombination rate. We also thank J. Edelman for assistance with the Scaffolding With DISCOVAR pipeline and J. Zhu and L. Nakhleh for their guidance with the PhyloNet toolkit. We thank E. Harney for assistance with qpGraph, as well as N. Rosser, F. Seixas, T. Xiong, P. Muralidhar and C. Veller for illuminating discussions.

Funding: This project was funded by a SPARC Grant from the Broad Institute of Harvard and MIT and startup and studentship funds from Harvard University for JM and NBE. Additional funding through NSF grant DGE1745303 to MM; BBSRC, Core Strategic Programme Grant BB/CSP17270/1 at the Earlham Institute for BC, GGA, and FDP; a Herchel Smith Postdoctoral Research Fellowship and a Smithsonian Tropical Research Institute Fellowship for JWD; NSF grant OIA 1736026 to BC and RP; Brazilian National Council for Scientific and Technological Development” (CNPq) grant 309853/2014–1 to GRPM; Fondos Concursables Universidad del Rosario 2016-P IN-2017–001 to CS; Marie Sklodowska-Curie fellowship (FITINV, N 655857) and NSERC fellowship for MC; NERC award UKSBS PR18037 to MB; NSF grant IOS-1656514 to RDR; NSF grant IOS-1656389 to RP; NERC NE/K012886/1 to KKD; NSF grant IOS-1452648 and NIH grant GM108626 to MK; MJ is supported by French National Research Agency: ANR-12-JSV7–0005-HybEvol and ANR-18-CE02–0019-Supergene. AB was supported by NIH grants 5U54CA193313 and GG010211-R01-HIV, and AFOSR grant FA9550–18-1–0415. DEN was supported in part by Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Grant Number U19AI110818 to the Broad Institute. DBJ’s work on *DISCOVAR de novo* was funded in part with federal funds from the National Human Genome Research Institute, US National Institutes of Health, US Department of Health and Human Services, under grants R01HG003474 and U54HG003067. LepBase development was supported by BBSRC grants BB/K020161/1 and BB/R015325/1.

References and Notes

1. Schluter D, The Ecology of Adaptive Radiation (OUP Oxford, 2000).
2. Lamichhaney S, Berglund J, Almen MS, Maqbool K, Grabherr M, Martinez-Barrio A, Promerová M, Rubin C-J, Wang C, Zamani N, Grant BR, Grant PR, Webster MT, Andersson L, Evolution of Darwin’s finches and their beaks revealed by genome sequencing. Nature 518, 371–375 (2015). [PubMed: 25686609]
3. Pease JB, Haak DC, Hahn MW, Moyle LC, Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol. 14, e1002379 (2016). [PubMed: 26871574]

4. Kozak KM, Wahlberg N, Neild AFE, Dasmahapatra KK, Mallet J, Jiggins CD, Multilocus species trees show the recent adaptive radiation of the mimetic *Heliconius* butterflies. Syst. Biol. 64, 505–524 (2015). [PubMed: 25634098]

5. Kozak KM, McMillan O, Joron M, Jiggins CD, Genome-wide admixture is common across the Heliconius radiation. bioRxiv, 414201 (2018).

6. Brower AVZ, Orduña IJG, Missing data, clade support and "reticulation": the molecular systematics of Heliconius and related genera (Lepidoptera: Nymphalidae) re-examined. Cladistics 34, 151–166 (2018).

7. Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, Blaxter M, Manica A, Mallet J, Jiggins CD, Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Research 23, 1817–1828 (2013). [PubMed: 24045163]

8. Martin SH, Davey JW, Salazar C, Jiggins CD, Recombination rate variation shapes barriers to introgression across butterfly genomes. PLoS Biol. 17, e2006288 (2019). [PubMed: 30730876]

9. Materials and methods are available as supplementary materials.

10. Drosophila 12 Genomes Consortium, Evolution of genes and genomes on the Drosophila phylogeny. Nature 450, 203–218 (2007). [PubMed: 17994087]

11. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D, Ancient admixture in human history. Genetics 192, 1065–1093 (2012). [PubMed: 22960212]

12. Wen D, Nakhleh L, Co-estimating reticulate phylogenies and gene trees from multi-locus sequence data. Syst. Biol. 61, 170–457 (2017).

13. Davey JW, Barker SL, Rastas PM, Pinharanda A, Martin SH, Durbin R, McMillan WO, Merrill RM, Jiggins CD, No evidence for maintenance of a sympatric *Heliconius* species barrier by chromosomal inversions. Evolution Letters 1, 138–154 (2017). [PubMed: 30283645]

14. Van Belleghem SM, Rastas P, Papanicolaou A, Martin SH, Arias CF, Supple MA, Hanly JJ, Mallet J, Lewis JJ, Hines HM, Ruiz M, Salazar C, Linares M, Moreira GRP, Jiggins CD, Counterman BA, McMillan WO, Papa R, Complex modular architecture around a simple toolkit of wing pattern genes. Nat. Ecol. Evol. 1, 52 (2017). [PubMed: 28523290]

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

16. Coyne JA, Orr HA, Speciation (Sinauer Associates Incorporated, 2004).

17. Begun DJ, Aquadro CF, Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356, 519–520 (1992). [PubMed: 1560824]

18. Schumer M, Xu C, Powell DL, Durvasula A, Skov L, Holland C, Blazier JC, Sankararaman S, Andolfatto P, Rosenthal GG, Przeworski M, Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science 360, 656–660 (2018). [PubMed: 29674434]

19. Davey JW, Chouteau M, Barker SL, Maroja L, Baxter SW, Simpson F, Joron M, Mallet J, Dasmahapatra KK, Jiggins CD, Major improvements to the Heliconius melpomene genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution. G3: Genes|Genomes|Genetics 6, 695–708 (2016). [PubMed: 26772750]

20. Dobzhansky T, Genetics and the Origin of Species (Columbia University Press, 1937).

21. Orr HA, Turelli M, Dominance and Haldane's rule. Genetics 143, 613–616 (1996). [PubMed: 8722810]

22. Jiggins CD, Linares M, Naisbit RE, Salazar C, Yang ZH, Mallet J, Sex-linked hybrid sterility in a butterfly. Evolution 55, 1631–1638 (2001). [PubMed: 11580022]

23. Naisbit RE, Jiggins CD, Linares M, Salazar C, Mallet J, Hybrid sterility, Haldane's rule and speciation in *Heliconius cydno* and *H. melpomene*. Genetics 161, 1517–1526 (2002). [PubMed: 12196397]

24. Van Belleghem SM, Baquero M, Papa R, Salazar C, McMillan WO, Counterman BA, Jiggins CD, Martin SH, Patterns of Z chromosome divergence among *Heliconius* species highlight the importance of historical demography. Mol. Ecol. 27, 3852–3872 (2018). [PubMed: 29569384]

25. Joron M, Papa R, Beltrán M, Chamberlain N, Mavárez J, Baxter S, Abanto M, Bermingham E, Humphray SJ, Rogers J, Beasley H, Barlow K, ffrench-Constant RH, Mallet J, McMillan WO, Jiggins CD, A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. PLoS Biol. 4, e303–10 (2006). [PubMed: 17002517]

26. Nadeau NJ, Pardo-Diaz C, Whibley A, Supple MA, Saenko SV, Wallbank RWR, Wu GC, Maroja L, Ferguson L, Hanly JJ, Hines H, Salazar C, Merrill RM, Dowling AJ, ffrench-Constant RH, Llaurens V, Joron M, McMillan WO, Jiggins CD, The gene *cortex* controls mimicry and crypsis in butterflies and moths. Nature 534, 106–110 (2016). [PubMed: 27251285]

27. Jay P, Whibley A, Frezal L, Rodríguez de Cara MÁ, Nowell RW, Mallet J, Dasmahapatra KK, Joron M, Supergene evolution triggered by the introgression of a chromosomal inversion. Curr. Biol. 28, 1839–1845.e3 (2018). [PubMed: 29804810]

28. Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, Whibley A, Becuwe M, Baxter SW, Ferguson L, Wilkinson PA, Salazar C, Davidson C, Clark R, Quail MA, Beasley H, Glithero R, Lloyd C, Sims S, Jones MC, Rogers J, Jiggins CD, ffrench-Constant RH, Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature 477, 203–206 (2011). [PubMed: 21841803]

29. Roda F, Mendes FK, Hahn MW, Hopkins R, Genomic evidence of gene flow during reinforcement in Texas *Phlox*. Mol. Ecol. 26, 2317–2330 (2017). [PubMed: 28141906]

30. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, Jiang X, Hall AB, Catteruccia F, Kakani E, Mitchell SN, Wu Y-C, Smith HA, Love RR, Lawniczak MK, Slotman MA, Emrich SJ, Hahn MW, Besansky NJ, Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347, 1258524–1258524 (2015). [PubMed: 25431491]

31. Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O, Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat. Commun. 8, 14363–11 (2017). [PubMed: 28186104]

32. Good JM, Vanderpool D, Keeble S, Bi K, Negligible nuclear introgression despite complete mitochondrial capture between two species of chipmunks. Evolution 69, 1961–1972 (2015). [PubMed: 26118639]

33. Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL, Speciation and introgression between Mimulus nasutus and Mimulus guttatus. PLoS Genet. 10, e1004410 (2014). [PubMed: 24967630]

34. Gante HF, Matschiner M, Malmstram M, Jakobsen KS, Jentoft S, Salzburger W, Genomics of speciation and introgression in Princess cichlid fishes from Lake Tanganyika. Mol. Ecol. 25, 6143–6161 (2016). [PubMed: 27452499]

35. Martin SH, Davey JW, Jiggins CD, Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol. Biol. Evol. 32, 244–257 (2015). [PubMed: 25246699]

36. The Heliconius Genome Consortium, Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487, 94–98 (2012). [PubMed: 22722851]

37. Broad Institute, DISCOVAR: Assemble genomes, find variants. https://www.broadinstitute.org/software/discovar/blog (2015).

38. Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ, Berger B, KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics 33, 574–576 (2017). [PubMed: 27797770]

39. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T-W, Wang J, SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1, 18 (2012). [PubMed: 23587118]

40. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC, GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33, 2202–2204 (2017). [PubMed: 28369201]

41. Margais G, Kingsford C, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27, 764–770 (2011). [PubMed: 21217122]

42. Smit A, Hubley R, Green P, RepeatMasker Open-4.0. www.repeatmasker.org (2013).

43. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212 (2015). [PubMed: 26059717]

44. Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D, Cactus: Algorithms for genome multiple sequence alignment. Genome Research 21, 1512–1528 (2011). [PubMed: 21665927]

45. Hickey G, Paten B, Earl D, Zerbino D, Haussler D, HAL: a hierarchical format for storing and anal yzing multiple genome alignments. Bioinformatics 29, 1341–1342 (2013). [PubMed: 23505295]

46. Hubisz MJ, Pollard KS, Siepel A, PHAST and RPHAST: phylogenetic analysis with space/time models. Brief. Bioinformatics 12, 41–51 (2011). [PubMed: 21278375]

47. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D, The human genome browser at UCSC. Genome Research 12, 996–1006 (2002). [PubMed: 12045153]

48. Misof B, Misof K, A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Syst. Biol. 58, 21–34 (2009). [PubMed: 20525566]

49. Kück P, ALICUT: a Perlscript which cuts ALISCORE identified RSS. Department of Bioinformatics, Zoologisches Forschungsmuseum A. Koenig (ZFMK), Bonn, Germany, version 2 (2009).

50. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS, ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589 (2017). [PubMed: 28481363]

51. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol 32, 268–274 (2015). [PubMed: 25371430]

52. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS, UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol 35, 518–522 (2018). [PubMed: 29077904]

53. Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T, ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30, i541–8 (2014). [PubMed: 25161245]

54. Kück P, Meusemann K, FASconCAT: Convenient handling of data matrices. Mol. Phylogenetics Evol 56, 1115–1118 (2010).

55. Quinlan AR, BEDTools: The Swiss-army tool for genome feature analysis. Curr. Protoc. Bioinformatics 47, 11.12.1–34 (2014).

56. Paradis E, Schliep K, ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 9, 532 (2018).

57. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR, SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb. Genom. 2, e000056 (2016). [PubMed: 28348851]

58. Gel B, Serra E, karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics 33, 3088–3090 (2017). [PubMed: 28575171]

59. Chan AH, Jenkins PA, Song YS, Genome-wide fine-scale recombination rate variation in Drosophila melanogaster. PLoS Genet. 8, e1003090 (2012). [PubMed: 23284288]

60. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Moonshine AL, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA, From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr. Protoc. Bioinformatics 43, 11.10.1–11.10.33 (2013). [PubMed: 25431634]

61. Browning BL, Browning SR, Genotype imputation with millions of reference samples. Am. J. Hum. Genet 98, 116–126 (2016). [PubMed: 26748515]

62. Hudson RR, 2002, Generating samples under a Wright-Fisher neutral model of genetic variation. Genome Biol. Evol 18, 337–338 (2002).

63. Kelleher J, Etheridge AM, McVean G, Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput. Biol 12, e1004842 (2016). [PubMed: 27145223]

64. Rambaut A, Grassly NC, Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci 13, 235–238 (1997). [PubMed: 9183526]

65. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol 59, 307–321 (2010). [PubMed: 20525638]

66. Thorvaldsdottir H, Robinson JT, Mesirov JP, Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinformatics 14, 178–192 (2013). [PubMed: 22517427]

67. Weisenfeld NI, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, Sogoloff B, Tabbaa D, Williams L, Russ C, Nusbaum C, Lander ES, Broad Institute DB Jaffe, Comprehensive variation discovery in single human genomes. Nat. Genet 46, 1350–1355 (2014). [PubMed: 25326702]

68. Love RR, Weisenfeld NI, Jaffe DB, Besansky NJ, Neafsey DE, Evaluation of DISCOVAR de novo using a mosquito sample for cost-effective short-read genome assembly. BMC Genomics 17, 187–10 (2016). [PubMed: 26944054]

69. Clavijo B, Garcia Accinelli G, Wright J, Heavens D, Barr K, Yanes L, Di Palma F, W2RAP: a pipeline for high quality, robust assemblies of large complex genomes from short read data. bioRxiv, 110999 (2017).

70. Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL, The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. GigaScience 6, 705 (2017).

71. Paajanen P, Kettleborough G, Lopez-Girona E, Giolai M, Heavens D, Baker D, Lister A, Cugliandolo F, Wilde G, Hein I, Macaulay I, Bryan GJ, Clark MD, A critical comparison of technologies for a plant genome sequencing project. GigaScience (2019), doi:10.1093/gigascience/giy163.

72. Goodwin S, McPherson JD, McCombie WR, Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics 17, 333–351 (2016).

73. Ray DA, Grimshaw JR, Halsey MK, Korstian JM, Osmanski AB, Sullivan KAM, Wolf KA, Reddy H, Foley N, Stevens RD, Knisbacher B, Levy O, Counterman B, Edelman NB, Mallet J, Simultaneous TE analysis of 19 Heliconiine butterflies yields novel insights into rapid TE-based genome diversification and multiple SINE births and deaths. Genome Biol. Evol. (2019).

74. Ahola V, Lehtonen R, Somervuo P, Salmela L, Koskinen P, Rastas P, Välimäki N, Paulin L, Kvist J, Wahlberg N, Tanskanen J, Hornett EA, Ferguson LC, Luo S, Cao Z, de Jong MA, Duplouy A, Smolander O-P, Vogel H, McCoy RC, Qian K, Chong WS, Zhang Q, Ahmad F, Haukka JK, Joshi A, Salojärvi J, Wheat CW, Grosse-Wilde E, Hughes D, Katainen R, Pitkänen E, Ylinen J, Waterhouse RM, Turunen M, Vähärautio A, Ojanen SP, Schulman AH, Taipale M, Lawson D, Ukkonen E, Mäkinen V, Goldsmith MR, Holm L, Auvinen P, Frilander MJ, Hanski I, The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. Nat. Commun. 5, 4737 (2014). [PubMed: 25189940]

75. Nowell RW, Elsworth B, Oostra V, Zwaan BJ, Wheat CW, Saastamoinen M, Saccheri IJ, Van't Hof AE, Wasik BR, Connahs H, Aslam ML, Kumar S, Challis RJ, Monteiro A, Brakefield PM, Blaxter M, A high-coverage draft genome of the mycalesine butterfly *Bicyclus anynana*. GigaScience 6, 1–7 (2017).

76. Zhan S, Zhang W, Niitepõld K, Hsu J, Haeger JF, Zalucki MP, Altizer S, de Roode JC, Reppert SM, Kronforst MR, The genetics of monarch butterfly migration and warning colouration. Nature 514, 317–321 (2014). [PubMed: 25274300]

77. 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, A genetic mechanism for female-limited Batesian mimicry in *Papilio* butterfly. Nat. Genet 47, 405–409 (2015). [PubMed: 25751626]

78. Cong Q, Borek D, Otwinowski Z, Grishin NV, Skipper genome sheds light on unique phenotypic traits and phylogeny. BMC Genomics 16, 639 (2015). [PubMed: 26311350]

79. The International Silkworm Genome Consortium, The genome of a lepidopteran model insect, the silkworm Bombyx mori. Insect Biochem. Mol. Biol 38, 1036–1045 (2008). [PubMed: 19121390]

80. You M, Yue Z, He W, Yang X, Yang G, Xie M, Zhan D, Baxter SW, Vasseur L, Gurr GM, Douglas CJ, 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 JW, Smith SM, 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, A heterozygous moth genome provides insights into herbivory and detoxification. Nat. Genet 45, 220–225 (2013). [PubMed: 23313953]

81. Bruen TC, Philippe H, Bryant D, A simple and robust statistical test for detecting the presence of recombination. Genetics 172, 2665–2681 (2006). [PubMed: 16489234]

82. Peter BM, Admixture, population structure, and F-statistics. Genetics 202, 1485–1501 (2016). [PubMed: 26857625]

83. Zairis S, Khiabanian H, Blumberg AJ, Rabadan R, Genomic data analysis in tree spaces. arxiv.org (2016).

84. Martin SH, Möst M, Palmer WJ, Salazar C, McMillan WO, Jiggins FM, Jiggins CD, Natural selection and genetic diversity in the butterfly *Heliconius melpomene*. Genetics 203, 525–541 (2016). [PubMed: 27017626]

85. Wakeley J, Coalescent Theory (Roberts and Co, 2008).

86. de Mendiburu F, Una herramienta de análisis estadísticopara la investigatión agrícola (Universidad Nacional de Ingenieria, Lima, 2009).

87. Pinharanda A, Martin SH, Barker SL, Davey JW, Jiggins CD, The comparative landscape of duplications in *Heliconius melpomene* and *Heliconius cydno*. Heredity 118, 78–87 (2017). [PubMed: 27925618]

88. Nakhleh L, A metric on the space of reduced phylogenetic networks. IEEE/ACM Trans Comput Biol Bioinform 7, 218–222 (2010). [PubMed: 20431142]

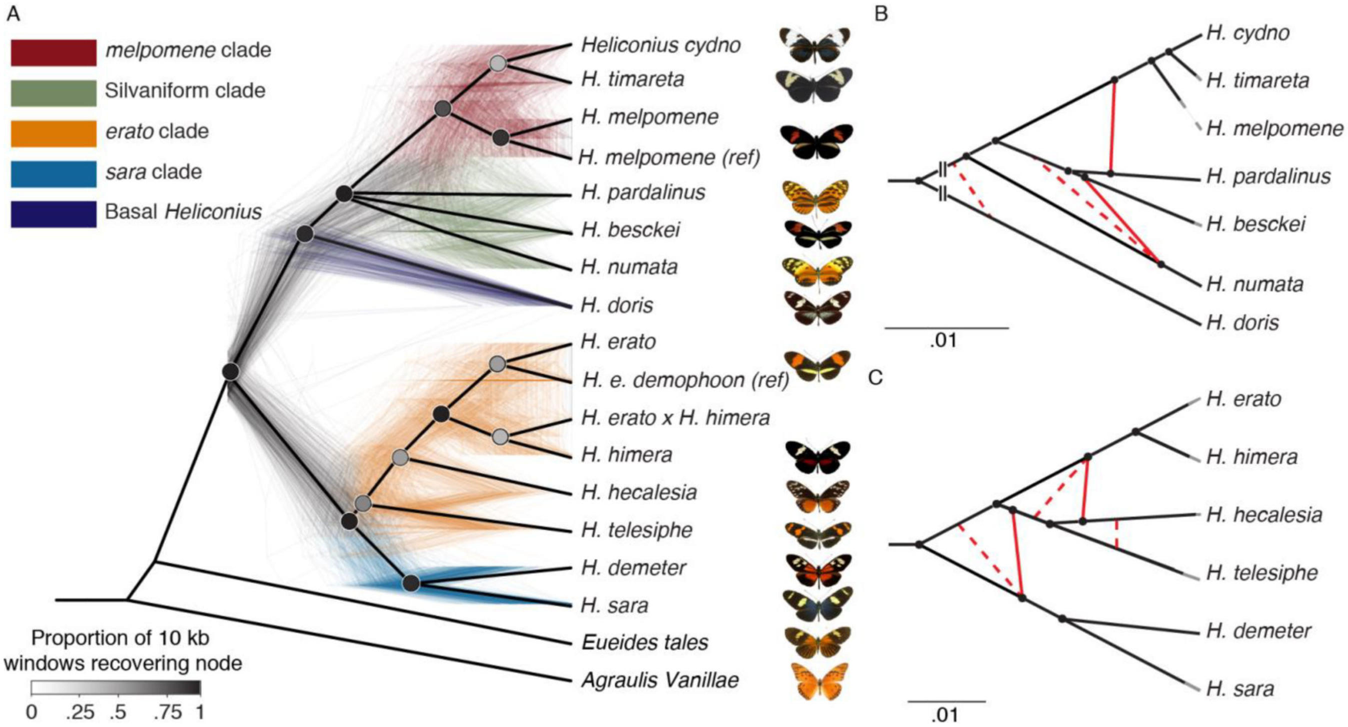

Fig. 1: Phylogeny and phylogenetic networks of *Heliconius* show lack of support for bifurcating tree.

A. All nodes resolved in a majority of species trees are shown in this cladogram (heavy black lines), while the poorly resolved silvaniform clade is collapsed as a polytomy (Fig. S20). The 500 colored trees were sampled from 10 kb non-overlapping windows and constructed with maximum likelihood. **B, C.** High-confidence tree structure (black) and introgression events (red) are shown as solid lines. Dashed red lines indicate weakly supported introgression events. Grey branch ends are cosmetic. The *melpomene*-silvaniform clade is shown in **B**, the *erato-sara* clade in **C.** Euclidean lengths of solid black lines are proportional to genetic distance along the branches. Scale bars in units of substitutions per site. Breaks at the base in **B** indicate that the branch leading to *H. doris* has been shortened for display.

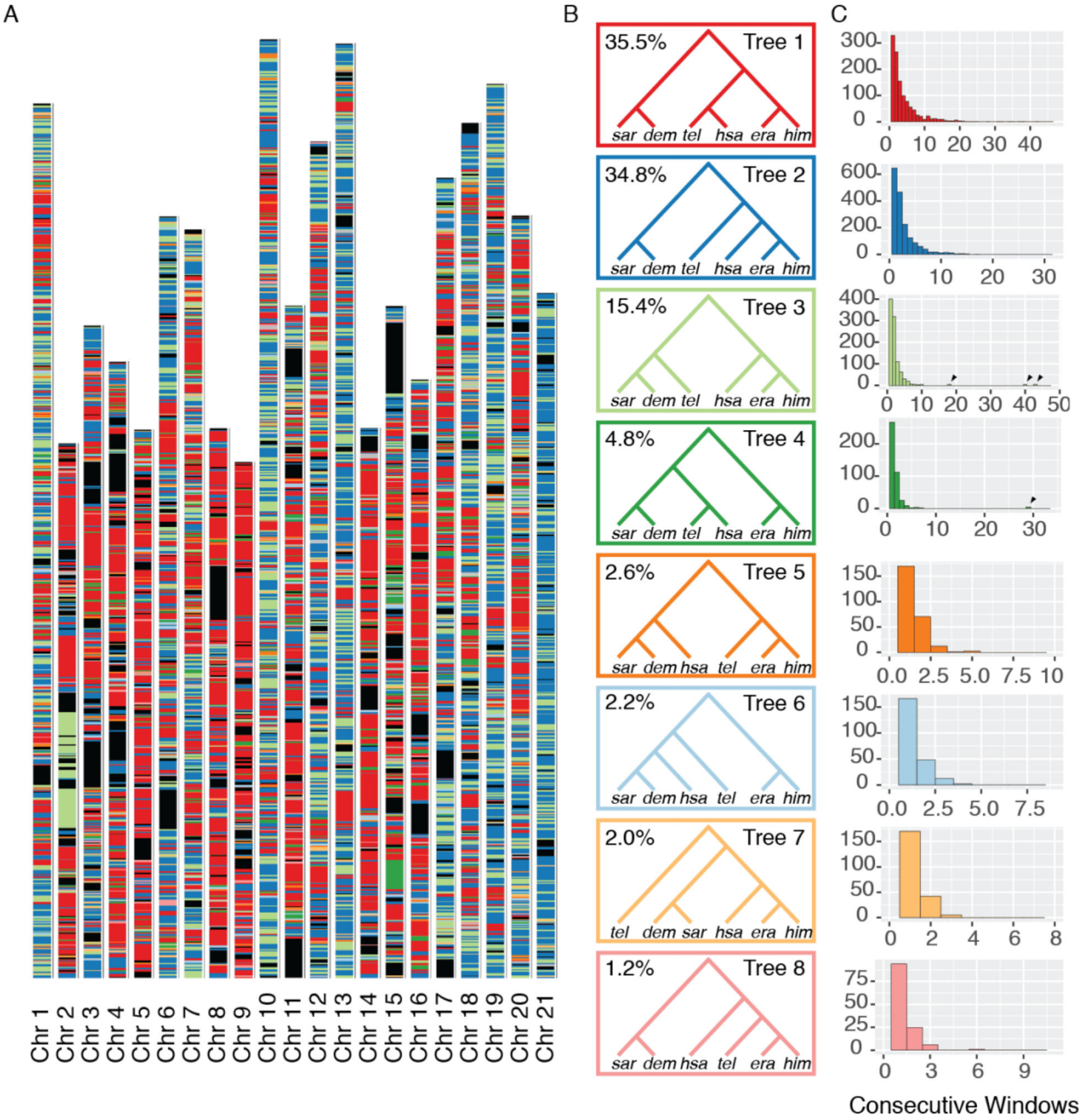

Fig. 2: Local evolutionary history in the *erato-sara* clade is heterogeneous across the genome.

A. Each bar represents a chromosome, in terms of the *H. erato* reference (14). Colored bands represent tree topologies of each 50 kb window; colors correspond to the topologies in **B**, with black regions showing missing data. **B.** The eight most common trees are shown. The value in the top left corner is the percentage of all 50 kb windows that recover that topology. **C.** Each histogram corresponds to the topology of the same color in **B**, and shows the distribution of the number of consecutive 50 kb windows with that topology. Arrows indicate long blocks in inversions.

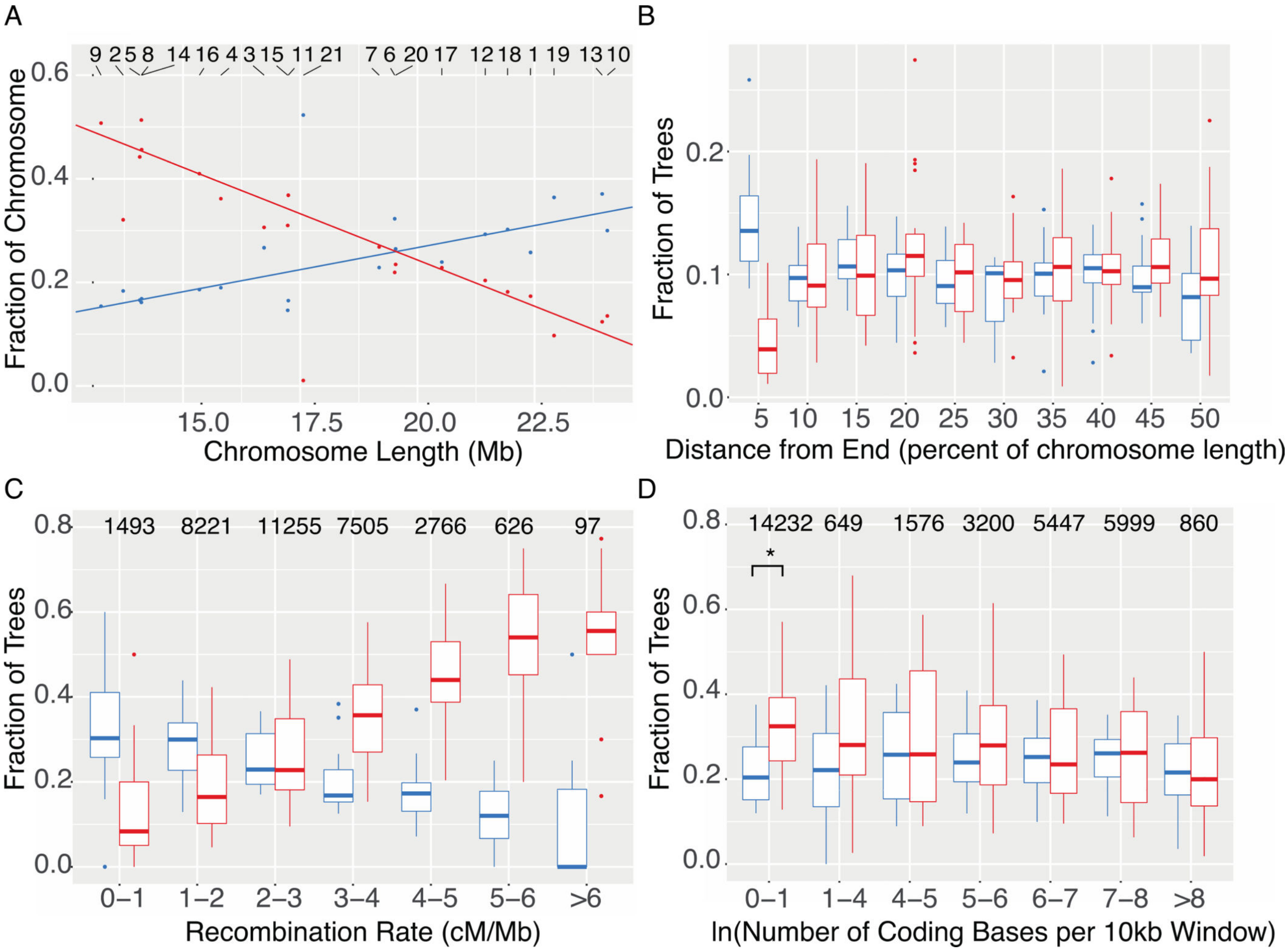

Fig. 3: Chromosomal architecture is strongly correlated with local topology.

Tree 1 is shown in red, and Tree 2 is shown in blue, as in Fig. 2. **A.** Tree 1 shows a negative relationship with chromosome size, while Tree 2 shows a positive relationship. Lines are linear regressions with chromosome 21 excluded. Numbers along top indicate chromosome number. **B.** Each chromosome was divided into 10 equally sized bins, and the occupancy of each topology in each bin was calculated as the number of windows that recovered the topology in the bin divided by the number of windows that recovered the topology in the chromosome. **C.** Windows are binned by recombination rate, and boxes show the fraction of each tree in each bin for each chromosome separately. Numbers above boxes are the number of windows in each bin. **D.** Boxes show the relationship of tree topology with coding density. Asterisk denotes significance at 5% level (paired t-test, p<0.025). In all boxplots, central line is median, box edges are first and third quartile, and whiskers extend to the largest value no further than 1.5*(inter-quartile range).

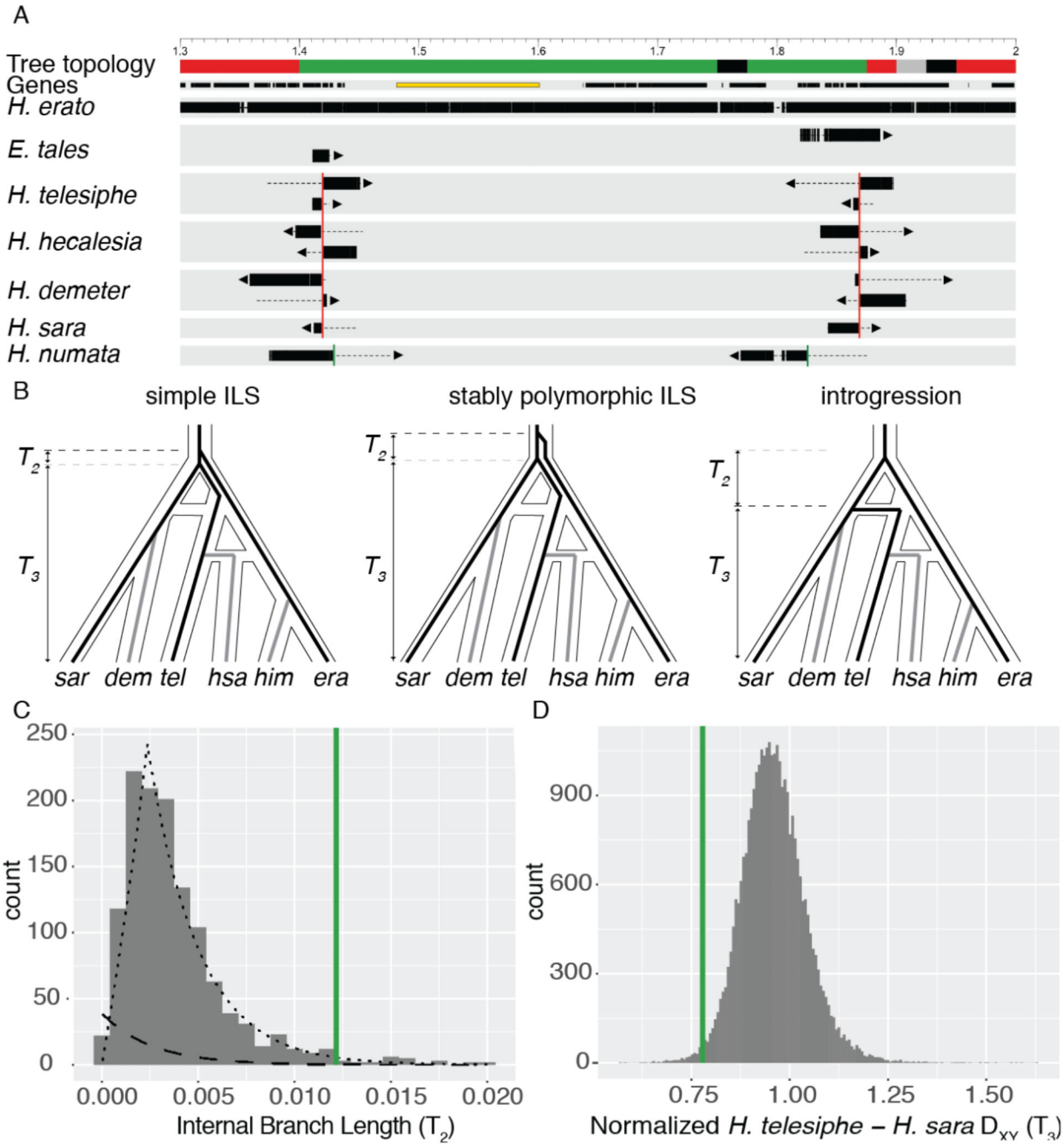

Fig. 4: Parallel evolution of a major inversion at the *cortex* supergene locus.

A. Map of 1.7 Mb region on chromosome 15. Coordinates are in terms of Hmel 2.5, and ticks are in Mb. Tree topology colors correspond to those in Fig. 2. Genes are shown as black rectangles; *cortex* is highlighted in yellow. Each line shows the mapping of a single contig. Aligned sections of each contig are shown as thick bars, while unaligned sections are shown as dotted lines. Arrows indicate the strand of the alignment. The *H. erato* group breakpoints are shown with red vertical lines, while the *H. numata* breakpoints are shown with green vertical lines. **B.** Evolutionary hypotheses consistent with the topology observed in this inversion in the context of the previously estimated phylogenetic network. The three species used in the triplet gene tree method – *H. erato, H. telesiphe,* and *H. sara* – are shown as black lines, while lineages not included are shown as grey lines. **C.** Histogram of internal branch lengths (T_2) in windows with the topology *H. erato, (H. telesiphe, H. sara)*. The inferred ILS distribution is shown as a dashed line, and the inferred introgression distribution is shown as a dotted line. The average internal branch length in the inversion is shown as a green vertical line. **D.** Histogram of normalized D_{XY} (T_3) between *H. telesiphe* and *H. sara.* Mean normalized D_{XY} in the inversion is shown as a green vertical line