Utah State University

DigitalCommons@USU

Ecology Center Publications

Ecology Center

5-1-2020

Recent Hybrids Recapitulate Ancient Hybrid Outcomes

Samridhi Chaturvedi
Utah State University

Lauren K. Lucas
Utah State University

C. Alex Buerkle
University of Wyoming

James A. Fordyce
University of Tennessee, Knoxville

Matthew L. Forister
University of Nevada, Reno

Chris C. Nice
Texas State University

See next page for additional authors

Follow this and additional works at: https://digitalcommons.usu.edu/eco_pubs

Part of the Ecology and Evolutionary Biology Commons

Recommended Citation

Chaturvedi, S., Lucas, L.K., Buerkle, C.A. et al. Recent hybrids recapitulate ancient hybrid outcomes. Nat Commun 11, 2179 (2020). https://doi.org/10.1038/s41467-020-15641-x

UtahState University
MERRILL-CAZIER LIBRARY

Authors

Samridhi Chaturvedi, Lauren K. Lucas, C. Alex Buerkle, James A. Fordyce, Matthew L. Forister, Chris C. Nice, and Zachariah Gompert

Check for updates

Recent hybrids recapitulate ancient hybrid outcomes

Samridhi Chaturvedi[1,2,3], Lauren K. Lucas[1], C. Alex Buerkle [4], James A. Fordyce[5], Matthew L. Forister[6], Chris C. Nice[7] & Zachariah Gompert[1,2 ✉]

Genomic outcomes of hybridization depend on selection and recombination in hybrids. Whether these processes have similar effects on hybrid genome composition in contemporary hybrid zones versus ancient hybrid lineages is unknown. Here we show that patterns of introgression in a contemporary hybrid zone in *Lycaeides* butterflies predict patterns of ancestry in geographically adjacent, older hybrid populations. We find a particularly striking lack of ancestry from one of the hybridizing taxa, *Lycaeides melissa*, on the Z chromosome in both the old and contemporary hybrids. The same pattern of reduced *L. melissa* ancestry on the Z chromosome is seen in two other ancient hybrid lineages. More generally, we find that patterns of ancestry in old or ancient hybrids are remarkably predictable from contemporary hybrids, which suggests selection and recombination affect hybrid genomes in a similar way across disparate time scales and during distinct stages of speciation and species breakdown.

[1] Department of Biology, Utah State University, Logan, UT 84322, USA. [2] Ecology Center, Utah State University, Logan, UT 84322, USA. [3] Department of Organismic & Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [4] Department of Botany, University of Wyoming, Laramie, WY 82071, USA. [5] Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN 37996, USA. [6] Department of Biology, University of Nevada, Reno, NV 89557, USA. [7] Department of Biology, Texas State University, San Marcos, TX 78666, USA. ✉email: zach.gompert@usu.edu

Numerous species hybridize[1,2] or show genomic evidence of ancient admixture[3–5]. Consequently, many organisms have genomes that are mosaics of chromosomal segments with ancestry from different lineages or species (e.g., refs. [6–11]). In hybrid zones, where gene flow and hybridization are ongoing, parental chromosomal segments are repeatedly introduced, and hybrids often vary substantially in genome composition[12–14]. In contrast, when gene flow and hybridization cease, as occurs during hybrid speciation or if one of the hybridizing taxa goes extinct, genome stabilization can occur whereby recombination causes ancestry segment size to decay and ancestry segments fix by drift or selection in the nascent hybrid lineage[7,15,16].

In both cases, the genomic mosaic of ancestry segments is shaped by recombination and selection, including selection against developmental incompatibilities and selection for locally adaptive ancestry combinations[12,14]. In contemporary hybrid zones with ongoing gene flow, selection results in differential or restricted introgression of some ancestry segments (e.g., barrier loci)[17–21], whereas selection causes shifts in the frequency of ancestry segments (including fixation) in old or ancient, stabilized, or partially stabilized hybrid populations[10,16,22]. Recombination interacts with selection to shape hybrid genomes[16,23–25]. In regions of higher recombination, neutral or adaptive foreign alleles can more readily disassociate from deleterious alleles allowing them to introgress, whereas in regions of low recombination, selection can prevent large blocks of ancestry from introgressing. Changes in genome composition and patterns of linkage disequilibrium (mediated by recombination) during the transition from a hybrid zone to stabilized hybrid lineage could alter selection on ancestry segments. Likewise, the primary sources of selection could change from, for example, selection against developmental incompatibilities to selection for novel allele combinations that enhance fitness and persistence in a new environment.

Unfortunately, comparisons of genome composition in contemporary versus old or ancient hybrids (i.e., those showing progress toward genome stabilization) are mostly lacking, especially for natural hybrids (for lab hybrids, see refs. [16,22]). Consistency between patterns of introgression in contemporary hybrid zones and the genomic mosaic of ancestry in ancient hybrids would suggest a major role for selection in determining hybrid genome composition, and would establish connections between early and late stages of speciation, especially speciation with gene flow and hybrid speciation. For example, consistent patterns would suggest that the same genes or gene regions that prevent the fusion of hybridizing species also experience selection during the origin of hybrid lineages or species. Such consistency could allow genes or genetic regions of general importance for adaptation and speciation to be identified (e.g., genes with environment-independent effects on fitness) and aid in interpreting patterns of ancestry in cases where ancient admixture occurred but where contemporary hybrids are lacking (e.g., Homo sapiens × H. neanderthalensis[8]). In contrast, a lack of consistency might imply that genomic outcomes of hybridization are highly context-dependent, and thus difficult to predict (e.g., ref. [26]).

Here, we take advantage of natural hybridization in North American Lycaeides butterflies to test the consistency of genome composition between a contemporary hybrid zone and multiple, old or ancient hybrid lineages that have progressed toward genome stabilization. In North America, Lycaeides consists of a complex of four nominal species of small blue butterflies and numerous partially stabilized (i.e., old or ancient) hybrid populations or lineages[6,27,28]. Partially stabilized hybrid populations in the central Rocky mountains and Jackson Hole (hereafter Jackson Hole Lycaeides) arose following hybridization between L. idas and L. melissa within the past 14,000 years (we refer to these as

ancient hybrids hereafter)[29]. Like L. idas (and unlike L. melissa), these Jackson Hole Lycaeides exhibit obligate diapause and are univoltine; they also use the same larval host plant (Astragalus miser) as nearby L. idas populations[29–31]. In contrast, partially stabilized hybrid lineages in the Sierra Nevada and Warner mountains of western North America occupy extreme alpine habitats not used by the parental species and exhibit novel, transgressive phenotypes, such as strong preference for an alpine endemic host plant (A. whitneyi) and a lack of egg adhesion to the host substrate[6,32]. We recently found evidence suggestive of a narrow (1–2 km), contemporary hybrid zone between L. melissa and Jackson Hole Lycaeides near the town of Dubois, Wyoming at the edge of the Rocky mountains[28]. Butterflies in the putative hybrid zone use feral (i.e., naturalized), roadside alfalfa (Medicago sativa) as their host plant (similar to some L. melissa populations), and are found in close proximity (along the road) to this non-native plant.

In this study, we first verify that the Dubois population constitutes a contemporary hybrid zone between L. melissa and Jackson Hole Lycaeides; the latter is an ancient, partially stabilized hybrid lineage derived from L. melissa and L. idas (see refs. [29,30] and the "Results" below). We then use a mixture of genotyping-by-sequencing (GBS) and whole-genome sequence data to compare patterns of introgression in this hybrid zone with the genomic mosaic of ancestry in Jackson Hole Lycaeides, and with additional ancient hybrid lineages in the Sierra Nevada and Warner mountains, and thereby quantify the consistency of hybrid genome composition across these cases. We expect less consistency with the Sierra Nevada and Warner mountains lineages as they differ in their ecology and in their taxonomic origin; nonetheless, these lineages represent genomic outcomes of hybridization between Lycaeides and are useful for understanding the general aspects of what transpires from initial hybridization, through isolation and stabilization. We show that patterns of ancestry in the ancient hybrids are remarkably predictable from contemporary hybrids, with a particularly striking lack of ancestry from Lycaeides melissa on the Z chromosome in both the ancient and contemporary hybrids. We thus use the contemporary Dubois hybrid zone as a window on the evolutionary process to show that similar processes operated during the origin and establishment of multiple, ancient hybrid lineages.

Results

Evidence of a contemporary hybrid zone in Dubois. Patterns of genetic variation across 39,193 SNPs and 23 populations (N = 835 butterflies) show that the Dubois, Wyoming population constitutes a contemporary hybrid zone between L. melissa and the ancient Jackson Hole hybrid lineage (Supplementary Table 1; Figs. 1, 2). Admixture proportions from entropy (ver. 1.2[28]) and a principal component analysis (PCA) of genetic variation show that Jackson Hole Lycaeides and the Dubois population are genetically intermediate between L. idas and L. melissa (Fig. 2b; Supplementary Figs. 1, 2). Jackson Hole Lycaeides form a relatively tight cluster in PCA space (especially individuals within populations) and show similar admixture proportions, consistent with past admixture but little ongoing gene flow from L. idas or L. melissa (Fig. 2b, c) (i.e., consistent with ancient rather than contemporary hybridization; see refs. [29,30], Supplementary Fig. 3 and 4 for additional support for this hypothesis, and details regarding geographic patterns of gene flow and genetic differentiation in Jackson Hole Lycaeides). In contrast, butterflies from the Dubois population span the entire genomic gradient from L. melissa to Jackson Hole Lycaeides (Fig. 2b, d). Thus, this single population, which occupies ~1–2 km along roadside alfalfa, exhibits greater variation in genome composition than the
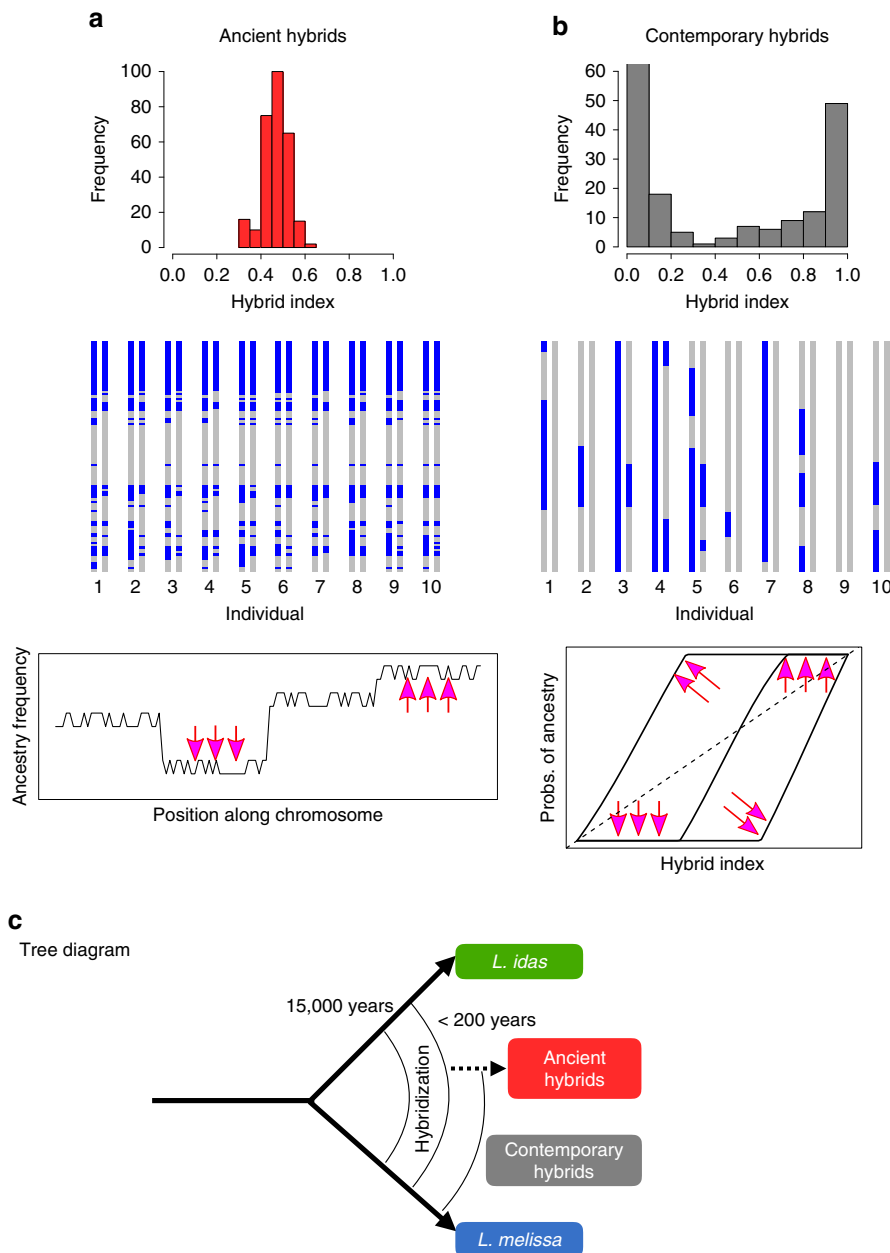
Fig. 1 Ancient versus contemporary hybrids. Conceptual overview and a comparative summary of genomic patterns expected in ancient hybrids (**a**) versus contemporary hybrids (**b**). Histograms show narrow (ancient) versus wide (contemporary) distributions of hybrid indexes. In ancient hybrids, ancestry blocks (gray vs. blue segments) have been broken up by recombination and some have stabilized, that is, fixed within the hybrid lineage. In contemporary hybrids, larger ancestry blocks are expected and these vary more among individuals. Plots show the expected effect of selection on ancestry frequencies and patterns of introgression in ancient and contemporary hybrids, respectively. In ancient hybrids, selection (pink arrows) shifts ancestry frequencies. In contemporary hybrids, selection shifts genomic clines for individual loci relative to the genome-wide average (dashed line). **c** Diagram represents the hypothesized history of hybridization in *Lycaeides*. Our results suggest that Jackson Hole *Lycaeides* are ancient hybrids, with ancestry blocks from *L. idas* and *L. melissa* (akin to panel **a**), whereas Dubois are contemporary hybrids with ancestry blocks from Jackson Hole *Lycaeides* and *L. melissa* (akin to panel **b** with colors denoting Jackson Hole vs. *L. melissa* ancestry).

entirety of Jackson Hole *Lycaeides*, with a range of more than 10,000 km² [29]. Also consistent with ongoing hybridization in the Dubois population, these butterflies exhibit elevated coupling (positive) linkage disequilibrium (Supplementary Figs. 5 and 6) and intermediate allele frequencies (Supplementary Figs. 7–12 for analogous results with Jackson Hole *Lycaeides*) at ancestry-informative SNPs (i.e., SNPs with an allele frequency difference of 0.3 or more between *L. idas* and *L. melissa*). Last, a discriminant analysis of genetic PCs confirmed that *L. melissa* and Jackson Hole *Lycaeides*, but not *L. idas*, occur in the Dubois hybrid zone

(Supplementary Fig. 13). Taken together, these results demonstrate ongoing hybridization in Dubois between *L. melissa* and nearby Jackson Hole *Lycaeides*, which are themselves a product of ancient hybridization between *L. idas* and *L. melissa* [29].

Genome composition in the ancient Jackson Hole hybrids. We next quantified genome-wide variation in the frequency of *L. idas* versus *L. melissa* ancestry segments in each of nine Jackson Hole *Lycaeides* populations (Supplementary Table 1). This includes a population adjacent to the Dubois hybrid zone (Bald Mountain
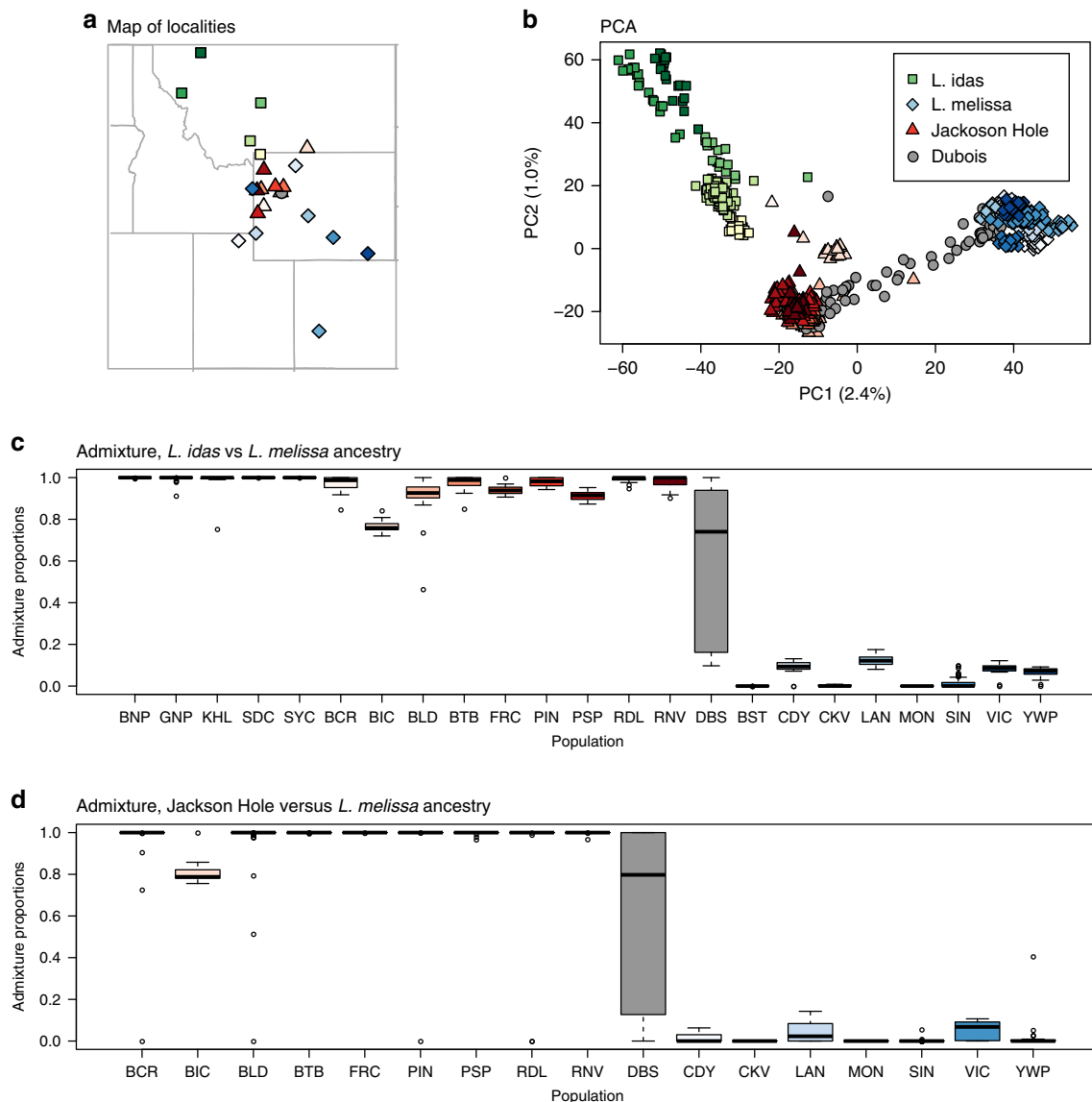
Fig. 2 Summary of population genetic patterns. a Map of sample locations with points of shapes based on nominal taxa and colored for different populations within taxa (Supplementary Table 1). **b** Ordination of genetic variation via principal component analysis (PCA). Points denote individuals (a few low-coverage individuals were removed for visualization). **c** Boxplots of admixture proportion estimates from entropy with $k = 2$ source populations for all populations included in the study ($n = 835$ butterflies from 23 populations). Boxes denote the 1st and 3rd quartile with the median given by the midline; whiskers extend to the minimum and maximum value or 1.5x the interquartile range with points for more extreme values. Tick marks below the plots identify populations based on the population abbreviations in Supplementary Table 1. **d** Boxplots of admixture proportion estimates from entropy with $k = 2$ source populations for all populations except *L. idas* ($N = 659$ butterflies from 18 populations), and boxes defined as in panel (**c**). Source data are provided as a Source Data file.

[BLD] that is ~5 km from Dubois), and much more distant populations (>100 km from Dubois, Fig. 2a). We focused this analysis on 1164 ancestry-informative markers (AIMs; SNPs with allele frequency differences between *L. idas* and *L. melissa* ≥0.3). We estimated ancestry segment frequencies using the correlated beta-process model implemented in popanc (ver. 0.1[15]). This method is similar to a hidden Markov model and accounts for the expected autocorrelation in ancestry along chromosomes, but allows ancestry frequencies to vary along the genome. Thus, it is suitable for quantifying ancestry frequencies in partially stabilized, old, or ancient hybrid lineages. The average frequency of *L. idas* ancestry varied from 0.63 (RDL) to 0.51 (BCR), resulting in a modest northwest–southeast cline in mean genome composition (consistent with ref. [29]) (Fig. 3). However, ancestry frequencies varied considerably within and among chromosomes, with

7.7–33.2% of the genome fixed or nearly fixed (i.e., frequency ≥0.95) for *L. melissa* or *L. idas* ancestry (Fig. 3; Supplementary Fig. 14). More of the genome was fixed (or nearly fixed) for *L. idas* ancestry than *L. melissa* ancestry (mean = 10.9% vs. 3.5%), and overall rates of fixation were higher on the Z chromosome than on autosomes (mean = 30.1% on the Z). Similar results were obtained when only male butterflies were analyzed (Supplementary Figs. 15 and 16).

Regions of exceptionally high *L. idas* ancestry frequencies were especially pronounced on the Z chromosome, and this held across all nine Jackson Hole populations (21.3–32.0% of the 225 AIMs on the Z chromosome had *L. idas* ancestry frequencies ≥0.95). Specifically, randomization tests showed that the 10% of AIMs with the highest *L. idas* ancestry frequencies were found on the Z chromosome 1.28–3.32 times more often than expected by chance
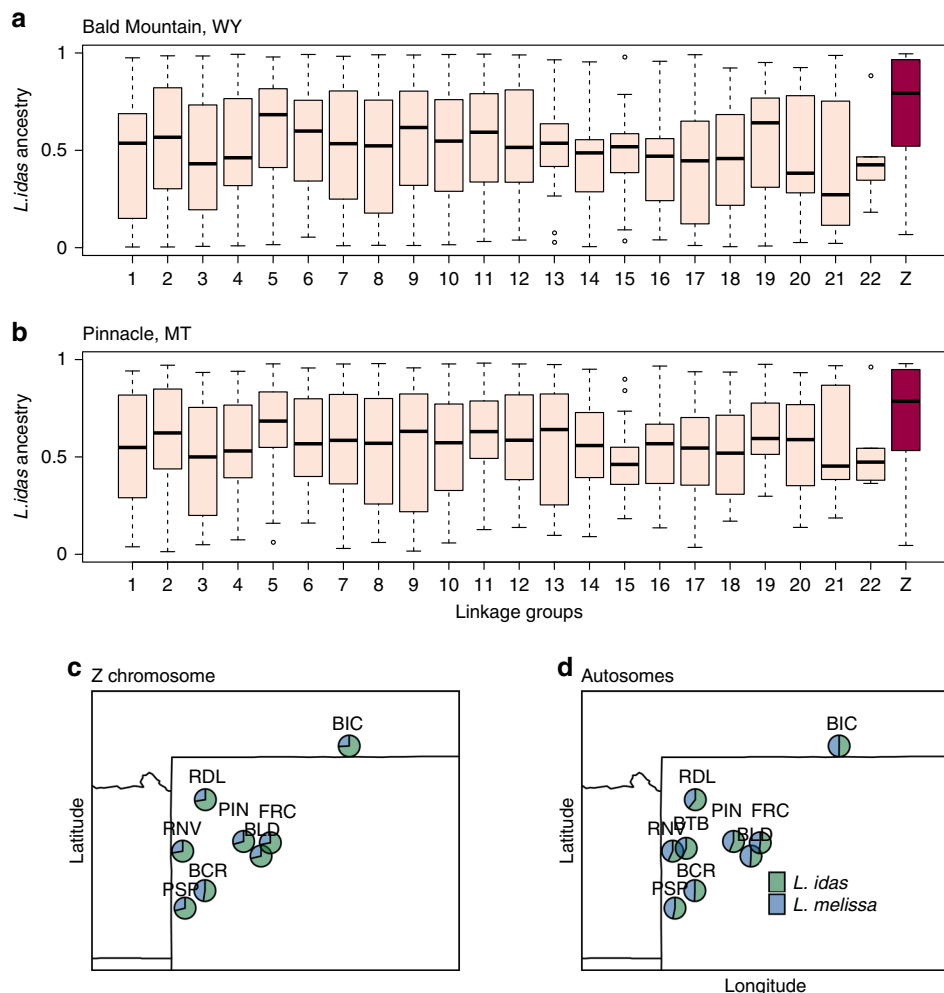
Fig. 3 Patterns of ancestry in the ancient hybrids. Boxplots show the distribution of *L. idas* ancestry across ancestry-informative SNPs (AIMs) for each linkage group in two representative Jackson Hole *Lycaeides* populations—Bald Mountain, WY (BLD, *n* = 74 butterflies) (**a**) and Pinnacle, WY (PIN, *n* = 20 butterflies) (**b**). See Supplementary Fig. 16 for additional populations. The Z-sex chromosome is shown in red. Boxes denote the 1st and 3rd quartile with the median given by the midline; whiskers extend to the minum and maximum value or 1.5x the interquartile range with points for more extreme values. Panels (**c**) and (**d**) show maps with pie charts reflecting the proportion of *L. idas* and *L. melissa* ancestry (mean) for the Z chromosome (**c**) and autosomes (**d**) for each of the nine populations (see Supplementary Table 1 for population IDs). Source data are provided as a Source Data file.

(Supplementary Table 2). In contrast, we found little evidence that regions of high *L. melissa* ancestry were overrepresented on the Z chromosome (Supplementary Table 3). Genomic regions with the highest levels of *L. idas* ancestry (i.e., the top 10% of AIMs with the highest *L. idas* ancestry) were in or near (within 1000 bp) genes (observed = 64, x-fold enrichment = 1.22, one-sided $P = 0.03$) and in or near gene-coding sequences (observed = 56, x-fold enrichment = 1.25, one-sided $P = 0.02$) more often than expected by chance (Supplementary Table 4). A similar pattern held for regions of high *L. melissa* ancestry (Supplementary Table 5). Finally, regions of majority *L. melissa* ancestry (>50%) were less common on larger chromosomes (Pearson $r = -0.5$, two-sided $P = 0.015$), which instead harbored a greater proportion of genetic regions fixed or nearly fixed for *L. idas* ancestry (>95%) (Pearson $r = 0.48$, two-sided $P = 0.021$) (this pattern was weaker when the Z chromosome was excluded; Supplementary Fig. 17). Such a pattern is expected when selection acts on many loci and against alleles from the minor parent (the one that contributes less to overall ancestry) and when recombination (per bp) is lower on larger chromosomes, as neutral minor parent alleles then have less opportunity to

recombine away from deleterious minor parent alleles on larger chromosomes[25].

Patterns of introgression in the Dubois hybrid zone. We then fit a Bayesian genomic cline model with bgc (ver. 1.04b[33]) to quantify genome-wide variability in introgression between *L. melissa* and Jackson Hole *Lycaeides* in the Dubois hybrid zone. This method estimates clines in ancestry for individual genetic loci (e.g., SNPs) along a genome-average admixture gradient[14,34]. As such, it can be applied when hybrid zones are confined to a single geographic locality, such as the Dubois population. Unlike the method used for the ancient Jackson Hole hybrids, the genomic cline method performs best when a wide range of hybrids with different genome compositions exist, as is the case for the Dubois hybrid zone. Whereas our analysis of the Jackson Hole populations defined ancestry with respect to *L. idas* or *L. melissa*, here we classify genomic regions as having been inherited from Jackson Hole or *L. melissa*. We once again focused on the 1164 AIMs (these SNPs showed appreciable allele frequency differences between the reference Jackson Hole and *L. melissa* populations used to define source population ancestry in this
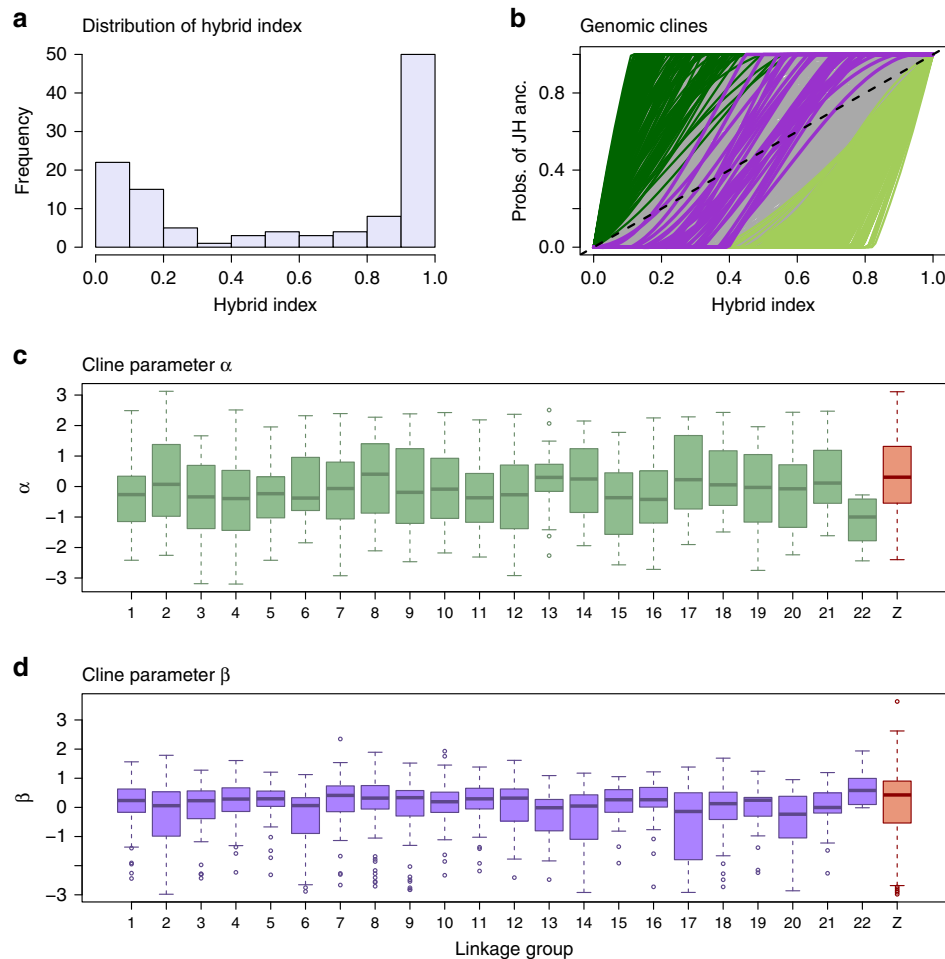
Fig. 4 Summary of the genomic cline analysis. a The histogram depicts the distribution of hybrid indexes in the Dubois hybrid zone. **b** This plot shows estimated genomic clines for a subset of ancestry-informative SNPs (AIMs). Each solid line gives the estimated probability of Jackson Hole (JH) ancestry for an AIM. Green lines denote cases of credible directional introgression (95% CIs for α that exclude zero) and purple lines denote credible cases of restricted introgression (95% CIs for $\beta > 0$) (gray lines denote clines not credibly different from the genome average). The dashed line gives the null expectation based on genome-wide admixture. Boxplots show the distribution of cline parameters α (**c**) and β (**d**) across loci for each linkage group (based on $n = 115$ butterflies). Boxes denote the 1st and 3rd quartile with the median given by the midline; whiskers extend to the minum and maximum value or 1.5× the interquartile range with points for more extreme values. Source data are provided as a Source Data file.

analysis, e.g., mean = 0.28, with differences greater than 0.1 for 78% of the AIMs). We detected credible variation in patterns of introgression across the genome (defined as cases where Bayesian 95% credible intervals (CIs) for genomic cline parameters did not span zero, Fig. 4). Of the 1164 AIMs, 34 (2.9%) showed credible evidence of restricted introgression (95% CIs for cline parameter $\beta > 0$) and constitute candidates for genomic regions harboring barrier loci between *L. melissa* and Jackson Hole *Lycaeides* (see ref. [35]; Fig. 4b). In total, 189 of the AIMs (16.2%) had credible evidence of excess Jackson Hole *Lycaeides* introgression relative to genome-average expectations (i.e., directional introgression of Jackson Hole alleles; 95% CIs for cline parameter $\alpha > 0$), and 273 AIMs (23.5%) had credible evidence of excess *L. melissa* introgression (i.e., directional introgression of *L. melissa* alleles; 95% CIs for cline parameter $\alpha < 0$). Estimates of α were mostly independent of the degree of difference in allele frequencies between Jackson Hole and *L. melissa* (e.g., Pearson $r = 0.007$; $r = -0.041$ for $|\alpha|$), but allele frequency differences explained 6.9% of the variation in point estimates of β and 20.9% of the variation in whether estimates of β were credibly greater than 0 (the latter likely reflects a combination of biological and statistical causes, e.g., high allele frequency differences likely reflect reduced introgression over time, but also result in increased power to

detect credible evidence of restricted introgression). Similar results were obtained when only male butterflies were analyzed; males carry two copies of the Z chromosome (Supplementary Fig. 18).

Genetic loci exhibiting directional or restricted introgression were distributed across the 23 *Lycaeides* chromosomes (Fig. 4b). However, randomization tests showed that AIMs with restricted introgression (34 loci with $\beta > 0$) and those with excess Jackson Hole introgression (189 loci with $\alpha > 0$) were found on the Z-sex chromosome more than expected by chance ($\beta > 0$, number on Z = 31, x-fold enrichment = 4.70, one-sided $P < 0.001$; $\alpha > 0$, number on Z = 67, x-fold enrichment = 1.84, one-sided $P < 0.001$) (Fig. 4c, d). This was not true for AIMs with directional introgression of *L. melissa* alleles (273 loci with $\alpha < 0$, number on Z = 37, x-fold enrichment = 0.70, $P = 0.998$). A greater proportion of AIMs exhibited restricted introgression on larger chromosomes (Pearson correlation between chromosome size and proportion of AIMs with $\beta > 0$, $r = 0.46$, two-sided $P = 0.03$), but this pattern was contingent on the Z chromosome, and similar patterns were not seen with respect to directional introgression AIMs (Supplementary Fig. 19). There was little evidence that genetic loci showing exceptional patterns of introgression in Dubois were clustered on genes or other
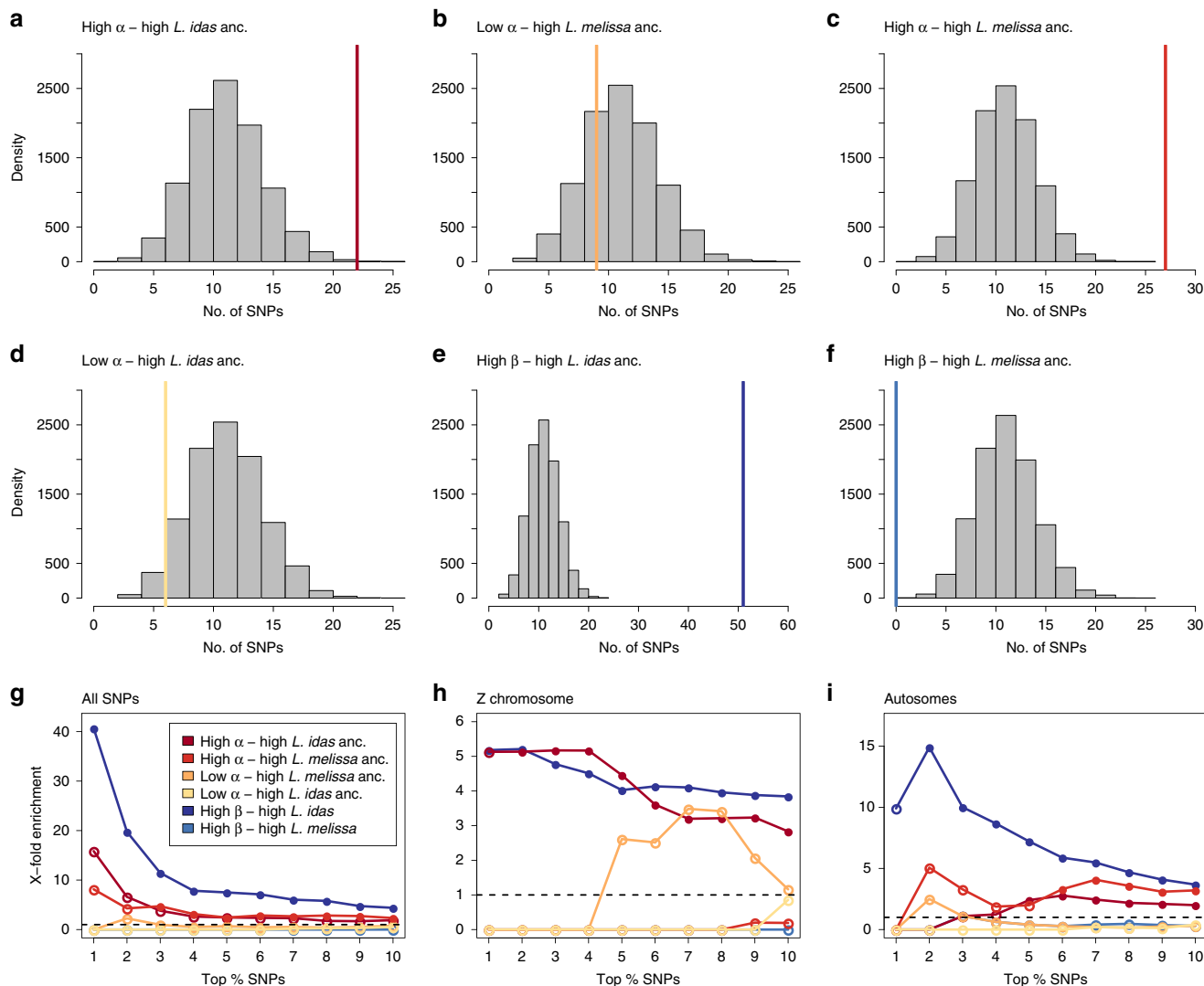
Fig. 5 Expected and observed numbers of SNPs with exceptional patterns of introgression in the Dubois hybrid zone and extreme ancestry frequencies in Jackson Hole *Lycaeides*. Panels **a–f** show results when considering the top 10% of AIMs in each category. Histograms give null expectations from randomization tests, and vertical solid lines show the observed number of AIMs exhibiting a given pattern. Comparisons shown are directional introgression of Jackson Hole alleles (high α) and high *L. idas* ancestry (in Jackson Hole) (**a**), directional introgression of *L. melissa* alleles (low α) and high *L. melissa* ancestry (**b**), directional introgression of Jackson Hole alleles (high α) and high *L. melissa* ancestry (**c**), directional introgression of *L. melissa* alleles (low α) and high *L. idas* ancestry (**d**), restricted introgression (high β) and high *L. idas* ancestry (**e**), and restricted introgression (high β) and high *L. melissa* ancestry (**f**). Panels **g–i** show how these results are affected by considering different levels of stringency (i.e., by examining the most extreme 10% to the top 1% of AIMs with each pattern), and when considering only the Z chromosome (**g**) or only the autosomes (**h**). Here, circles denote the ratio of the observed to expected overlap from the null, and the circles are filled ($P \leq 0.05$) or not ($P > 0.05$) to denote whether the overlap is greater than expected by chance from a one-sided randomization test. Source data are provided as a Source Data file.

structural genomic features (we did find some evidence of AIMs with credible *L. melissa* introgression being overrepresented in coding sequences and proteins; Supplementary Tables 6–8).

Genomic outcomes of hybridization are predictable. Regions of the genome showing exceptional introgression in the contemporary Dubois hybrid zone coincided with genomic regions exhibiting extreme ancestry frequencies in the ancient Jackson Hole hybrid lineage to a much greater extent than expected by chance (Fig. 5). For example, AIMs exhibiting restricted introgression in the Dubois hybrid zone (top 10% of AIMs with the highest β values from bgc) overlapped with those with the highest *L. idas* ancestry frequencies in Jackson Hole (top 10%) about four times more often than expected by chance (observed = 51, x-fold enrichment = 4.41, one-sided $P < 0.001$, Fig. 5e). The degree of

overlap increased when we considered more extreme cutoffs for the comparison, up to the top 1% of AIMs with the greatest degree of restricted introgression and the highest *L. idas* ancestry frequencies (x-fold enrichments ranged from 4.41 to 42.48, Fig. 5g). Similar patterns were observed when comparing AIMs with evidence of excess directional introgression of Jackson Hole alleles in Dubois (top 10% with the highest values of α) and those with the highest (i) *L. idas* (observed = 22, x-fold enrichment = 1.91, one-sided $P = 0.0014$) or (ii) *L. melissa* (observed = 26, x-fold enrichment = 2.24, one-sided $P < 0.001$) ancestry frequencies in Jackson Hole *Lycaeides* (Fig. 5a, c). The strength of this signal of excess overlap was again more pronounced when considering more extreme cutoffs up to the top 1% of AIMs (x-fold enrichment ranged from 1.91 to 16.93, Fig. 5f). In contrast, we found no evidence that AIMs with excess directional introgression of

L. melissa alleles in Dubois (top 10% with the lowest values of α) coincided with those AIMs with the highest *L. idas* or *L. melissa* ancestry frequencies in Jackson Hole *Lycaeides*, or evidence of excess overlap between AIMs exhibiting restricted introgression in Dubois and those with the highest *L. melissa* ancestry in Jackson Hole (Fig. 5b, d, f).

Evidence of consistency (i.e., predictability) of genome composition between contemporary Dubois hybrids and the ancient Jackson Hole hybrids comes from both the autosomes and Z-sex chromosome. Specifically, when we repeated the above comparisons with either the 22 autosomes or the Z chromosome alone, we obtained mostly similar results. For example, AIMs showing restricted introgression (top 10% with the highest values for β) and those with high *L. idas* ancestry frequencies in Jackson Hole *Lycaeides* (i.e., closest to being fixed for *L. idas* ancestry) coincided 3.7 times (observed = 37, one-sided $P < 0.0001$) and 3.8 times (observed = 38, one-sided $P < 0.0001$) more often than expected by chance for the autosomes and Z chromosome, respectively (Fig. 5h, i). Similarly, AIMs showing evidence of excess directional introgression of Jackson Hole alleles in Dubois (top 10%) and the highest frequencies of *L. idas* ancestry in the Jackson Hole populations coincided 1.9 (autosomes, observed = 22, one-sided $P = 0.0018$) and 2.8 (Z, observed = 12, one-sided $P < 0.0001$) times more often than expected by chance (Fig. 5g, h). We obtained similar results when defining AIMs as SNPs with an allele frequency difference of greater than 0.2 rather than 0.3 (2126 SNPs) (Supplementary Fig. 20), and when basing our analyses only on male butterflies, which carry two copies of the Z chromosome (Supplementary Figs. 21 and 22). Consistent results were also obtained when considering AIMs with the highest *L. idas* or *L. melissa* ancestry together (i.e., closest to being fixed for *L. idas* or *L. melissa* ancestry; Supplementary Fig. 23).

We next analyzed the potential functional significance of genetic regions harboring the AIMs that were in the top 10% for both restricted introgression in Dubois (high β) and high *L. idas* ancestry frequencies in the ancient Jackson Hole *Lycaeides* lineage. We focus on this set of 51 loci as we think these are our best candidates for tagging regions of the genome-harboring barrier loci, that is, regions of the genome that constitute the basis of (partial) reproductive isolation among these lineages. This assertion stems from the fact that these loci exhibited restricted introgression in the contemporary hybrid zone combined with extreme ancestry frequencies (putatively arising from selection) in the Jackson Hole populations (specifically, high *L. idas* ancestry frequencies as we did not detect excess overlap between AIMs exhibiting restricted introgression and those with high *L. melissa* ancestry frequencies). Twenty nine of these 51 AIMs were in or near genes (i.e., within 1000 bps of annotated gene boundaries) (in general, these barrier loci were not overrepresented in or near specific structural features of the genome; Supplementary Table 9, see also Supplementary Tables 10 and 11). These genes exhibited a range of predicted functions (Supplementary Tables 12–14), with several genes standing out as being of particular interest. This includes the Z-linked 6-phosphogluconate dehyrdogenase gene, *6-pgd* (Supplementary Fig. 24), which is associated with cold hardiness and diapause in other insects[36–38], and constitutes or is linked to a barrier locus in swallowtail butterflies[39–41]. Four AIMs were in or near two immunoglobulin superfamily genes; these were also on the Z chromosome (Supplementary Table 14 and Supplementary Fig. 24). This gene superfamily is crucial for pathogen defense in insects[42], and is associated with reproductive isolation in mice[43]. Also among this set of genes was an autosomal olfactory receptor/odorant-binding gene (Supplementary Fig. 24). Such genes are known to affect host plant use in other butterflies[44,45]. One of the 51 AIMs was within the autosomal nuclear pore gene *Nup93* (Supplementary Table 13).

This is part of the nuclear pore complex, which is involved in multiple Dobzhansky–Muller incompatabilities in *Drosophila*. Finally, two AIMs in *armadillo* or *armadillo*-like genes (one autosomal and one Z-linked) that are involved in the Wnt-signaling pathway and affect wing development in *Heliconius*, and other butterflies were among this set of candidate barrier loci[46]. A diversity of functions were also predicted for the subset of AIMs with both high directional introgression of Jackson Hole alleles in Dubois and high *L. idas* ancestry in Jackson Hole *Lycaeides*; this includes immunoglobulin superfamily genes (Supplementary Tables 15 and 16) (also see Supplementary Tables 17 and 18).

Extending predictability to additional hybrid lineages. We next asked whether patterns of introgression in the contemporary, Dubois hybrid zone were also predictive of patterns of ancestry in two additional, ancient hybrid lineages from the Sierra Nevada and Warner mountains of western North America[6,28,32]. These ancient hybrid lineages occupy alpine habitat on isolated mountains ~900–1000 km from Dubois. Past work suggests that these ecologically distinct populations arose within the past 2 million years (and perhaps more recently) following hybridization between *L. anna* and *L. melissa* (with a possible contribution from *L. idas* or a close relative of *L. idas/L. melissa*)[6,28,32,47]. Using whole-genome sequences from *L. anna*, *L. melissa*, *L. idas*, and the Sierra Nevada and Warner mountain lineages, we estimated phylogenetic relationships for the entire genome and for 1000 SNP windows along each chromosome to characterize patterns of ancestry in the ancient hybrids. Based on the results from Dubois and the Jackson Hole populations, we predicted reduced *L. melissa* ancestry on the Z chromosome, and topologies suggesting reduced introgression for the candidate barrier loci (i.e., regions containing the 51 AIMs with restricted introgression in Dubois and high *L. idas* ancestry frequencies in Jackson Hole).

Consistent with past work, the whole-genome consensus or species tree showed that the Warner mountain population was intermediate between *L. anna* and *L. melissa/L. idas*, whereas the Sierra Nevadan population was genetically more similar to *L. anna* (Fig. 6). Nonetheless, trees based on 1000 SNP windows varied across the genome (Supplementary Fig. 25). The species tree was the most common topology overall (29.9%), but several alternative trees were also common, especially on the autosomes (Fig. 6; Supplementary Figs. 26 and 27). Whereas the trees that differ from the species tree could reflect incomplete lineage sorting, tree topologies were autocorrelated along the genome, which suggests that introgression has contributed to these patterns as well (Supplementary Fig. 27).

The species (i.e., consensus) tree was seen about 30% of the time for autosomal windows, Z-chromosome windows, and the 1000 SNP windows that contained the 51 candidate barrier loci (Supplementary Fig. 26). The second most common autosomal tree suggested introgression from *L. melissa* into the Warner mountain lineage (24.6% of topologies). As predicted, this tree was significantly and substantially underrepresented on the Z chromosome (2.3%, x-fold = 0.10, one-sided $P < 0.001$) and for the candidate barrier loci (6.1%, x-fold = 0.43, one-sided $P = 0.047$) (Fig. 6). In contrast, a tree uniting the two ancient hybrid lineages was rare on the autosomes (4.3%), but commonly observed on the Z chromosome (26.2%, x-fold = 5.36, one-sided $P < 0.001$) and for the barrier loci (28.6%, x-fold = 3.62, one-sided $P = 0.008$) (Fig. 6). This suggests an ancient shared ancestry between these hybrid lineages that has been especially resistant to gene flow, or alternatively, Z-biased introgression between them. Finally, the introgression tree uniting the Warner mountain lineage and *L. melissa* was recovered more often on smaller
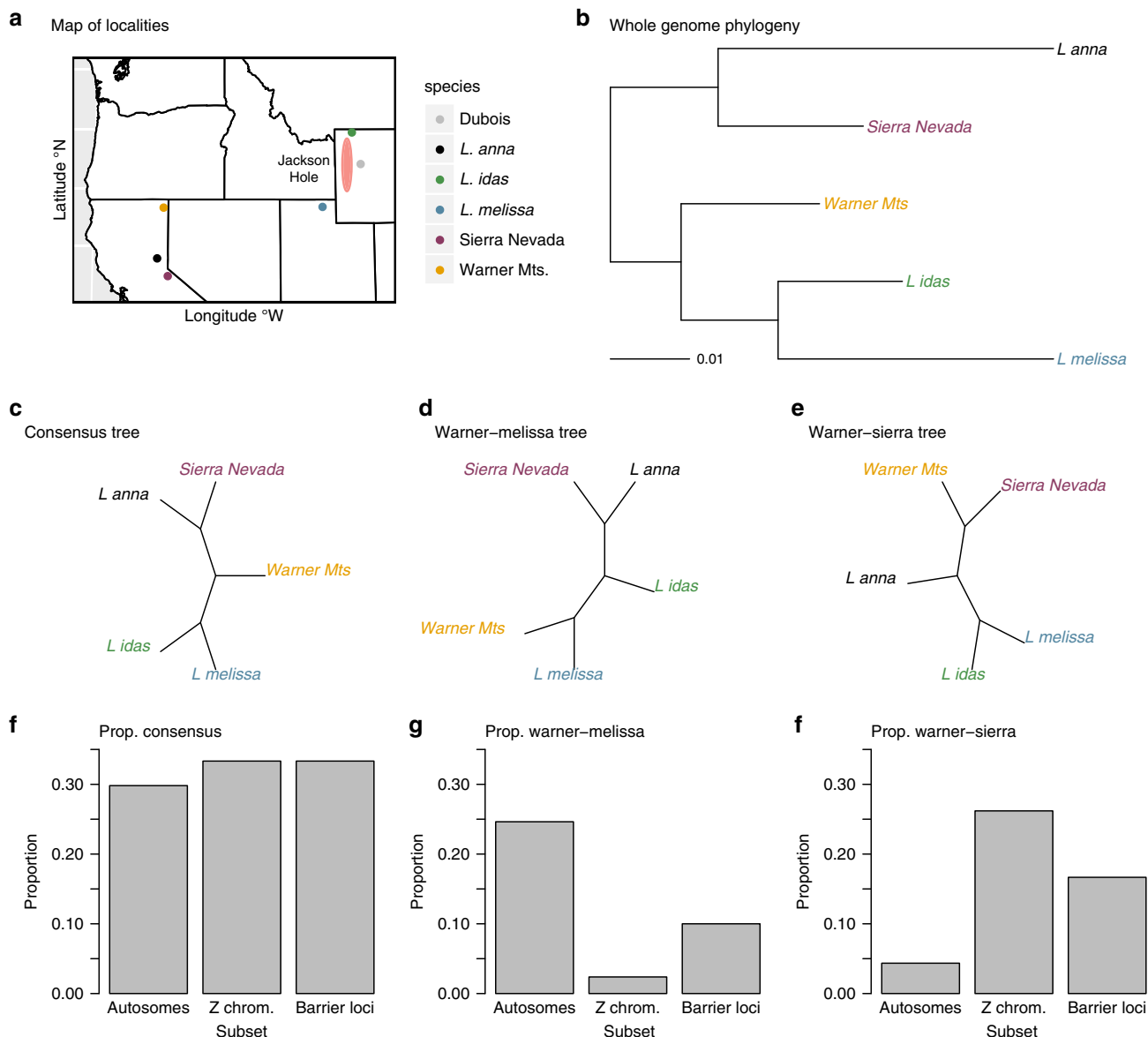
Fig. 6 Whole-genome phylogenetic analyses. Panel (**a**) shows a map of the sampling localities in the western United States; Dubois is included for reference but was not sampled for whole-genome phylogenetics. The Jackson Hole ancestry cline (i.e., the range of the ancient Jackson Hole hybrids) is shown for reference as well (red zone). Panel **b** shows the midpoint-rooted maximum likelihood phylogram inferred from the whole-genome SNP set (2,013,201 autosomal SNPs). Panels (**c**–**e**) show three common unrooted tree topologies inferred from 1000 SNP windows. The tree in panel (**c**) is the most common topology and matches the whole-genome tree in **b**. Trees in panels **d** and **e** differ by grouping the Warner Mts. population with *L. melissa* or the Sierra Nevada population, respectively. Panels (**f**–**h**) show the proportion of 1000 SNP topologies matching the trees in **c**–**e** for autosomal windows, Z-chromosome windows, and the 1000 SNP windows that contain the candidate barrier loci (see main text). There is a significant deficit of the topology in **d** on the Z chromosome and among the barrier loci (x-fold = 0.10, one-sided $P < 0.001$, and x-fold = 0.43, one-sided $P = 0.047$, respectively), and a significant excess of topology (**e**) on the Z chromosome and among the barrier loci (x-fold = 5.36, $P < 0.001$, and x-fold = 3.62, $P = 0.008$). See Supplementary Fig. 25 for the full set of tree topologies recovered. Source data are provided as a Source Data file.

chromosomes (Pearson correlation with LG size, $r = -0.68$, two-sided $P < 0.001$; autosomes only, $r = -0.61$, two-sided $P = 0.003$), whereas the tree uniting the ancient hybrid lineages was more common on large chromosomes (Pearson correlation with LG size, $r = 0.55$, two-sided $P = 0.006$; autosomes only, $r = 0.46$, two-sided $P = 0.033$) (Supplementary Fig. 28).

To explicitly test that the above patterns were due to differential introgression, we calculated admixture proportions (f_d) for each LG in the Warner mountain and Sierra Nevada lineages[48]. We found reduced introgression on the Z in both lineages (Sierra Nevada, one-sided $P = 0.043$; Warner mountain, one-sided $P < 0.001$; Supplementary Fig. 29A, B), and evidence of

a negative association between admixture and chromosome size in the Sierra Nevada lineage ($\beta = 1.6e^{-9}$, std. error = $6.4e^{-10}$, two-sided $P = 0.02$; Supplementary Fig. 29C, D). Thus, we find evidence of patterns of introgression and ancestry in these ancient hybrids consistent with patterns observed in the Dubois hybrid zone and ancient Jackson Hole hybrid lineage. Because the majority of the candidate barrier loci were Z-linked, the signal for these loci is not independent of the signal for the Z chromosome.

Discussion

Biologists have endeavored to connect microevolutionary processes to macroevolutionary changes that occur over longer

periods of time (e.g., refs. [49–51]). Here, we illustrate one way in which this gap can be bridged by showing that microevolutionary patterns and processes in a contemporary hybrid zone predict the genome composition of ancient (partially) stabilized hybrid taxa (a macroevolutionary-scale outcome). This mirrors past work in *Helianthus* showing that the genome composition of ancient hybrid sunflower species could be predicted from evolution in synthetic hybrids and QTL mapping experiments[22,52]. However, our results are novel in that they demonstrate that the genomic outcomes of hybridization can be predicted from natural hybrids as well, and despite differences in the ecological and genomic context of the different hybridization events. Additional work in other organisms where ancient and contemporary hybrids are found is needed to determine the generality of this result (e.g., *Populus* and *Helianthus*[53–55]). Beyond this, genomic analyses of multiple lineages at various points along the continuum from hybrid zone to stabilized hybrid species could prove particularly interesting. Such analyses could help distinguish between consistency arising from rapid fixation of ancestry segments versus sustained, consistent selection pressures over long periods of time. More generally, our results suggest that genomic analyses of both hybrid zones and lineages or species might offer a particularly tractable framework for assessing the ways in which and degree to which speciation (especially hybrid speciation) might be predictable from microevolutionary processes of selection and recombination.

Patterns of parallel or repeatable evolution strongly suggest a major role for the deterministic process of natural selection[56,57]. Thus, genes and gene regions with restricted introgression in Dubois, and extreme ancestry frequencies or phylogenetic relationships in Jackson Hole and the other ancient hybrid lineages represent strong candidates for harboring barrier loci (i.e., speciation genes). Given the differences in environment and ecological context across these instances of admixture, our results suggest that many of these putative barrier loci operate in a manner that is not highly environment dependent, which is consistent with a major role of intrinsic incompatibilities in the speciation process[58,59]. Predictability (i.e., consistency) was the highest when considering the Z chromosome, where we detected a lack of *L. melissa* ancestry across the hybrid zone and lineages. This could be explained by an increased efficacy of selection against Z loci in hemizygous individuals, or linkage and a higher density of selected loci on the Z chromosome in general[60,61]. An excess of intrinsic incompatibilities arising from epistatic interactions is an especially likely explanation, as the hybrid populations generally harbored less *L. melissa* ancestry (lower *L. melissa* ancestry could reflect selection or demographic aspects of the hybridization process). Nonetheless, our ability to predict the genome composition of ancient hybrids, especially the Jackson Hole lineage, held even when excluding the Z chromosome. Thus, while our results are consistent with theory and other empirical studies that suggest a disproportionate role for sex chromosomes in speciation (e.g., refs. [14,21,59,60]), this was not the sole factor that made genome composition predictable in this system. Indeed, our analyses highlighted candidate barrier loci on autosomes (e.g., an autosomal olfactory receptor/odorant- binding gene and *Nup93*), as well as on the Z chromosome. Our results also suggest that reduced introgression often occurs on longer chromosomes, which should undergo less recombination per base pair, in *Lycaeides*. This is consistent with selection acting on many loci scattered across the genome, and has been shown in several systems (e.g., refs. [16,23,25]). In our case, this pattern was driven in part by the Z chromosome, which is one of the largest chromosomes, but which should not otherwise experience reduced recombination (i.e., recombination in butterflies is completely suppressed in the heterogametic sex, including recombination

between homologous autosomes). We do not yet know whether similar results hold based on finer-scale variation in recombination rate within chromosomes, driven for example by structural variants or other regions of reduced recombination. However, our results do show that this pattern holds early and late in the hybridization and genome stabilization process, and that selection and recombination interact during this process in a manner that results in consistent or repeatable patterns of hybrid genome composition.

In conclusion, the fact that putative barrier loci in the Dubois hybrid zone exhibit extreme ancestry patterns in ancient hybrid lineages is consistent with the hypothesis that the same genes or gene regions that prevent species fusion during hybridization experience selection during the evolution of hybrid lineages or species. This hypothesis is also supported by studies in other systems showing that incompatibility loci contribute to hybrid speciation[16,52], and is generally consistent with accumulating data suggesting that the same genes are repeatedly involved in adaptation (e.g., refs. [62,63]). Thus, seemingly distinct facets of speciation, such as the maintenance of taxonomic boundaries across hybrid zones and the origin of novel biodiversity through admixture, may have a predictable, common genetic basis.

Methods

Genome assembly and annotation. We generated a new, chromosome-scale reference genome for *L. melissa* from information on proximity ligation of DNA in chromatin and reconstituted chromatin. Our previous genome assembly comprised 14,029 scaffolds (total assembly length = 360 megabase pairs [mbps]; scaffold N50 = 65 kilobase pairs (kbps)), which had been joined in a linkage map with 23 linkage groups (*Lycaeides* has 22 autosomes and a ZW sex chromosome system)[64,65]. In this study, we improved upon this assembly using DNA-sequence data from Chicago and Hi-C libraries[66,67]. Creation and sequencing of the Chicago and Hi-C libraries were outsourced to Dovetail Genomics. The new sequence data and our old assembly were combined using the HiRise assembler (also outsourced to Dovetail Genomics). The new *L. melissa* genome assembly has a final N50 of 15.5 mbps, and 90% of the genome comprises only 21 scaffolds. A whole-genome comparative alignment with mummer (version 3.2 with the maximal unique match setting[68]) of this new genome assembly with the previously published genome assembly and linkage map showed that each of our previously defined linkage groups corresponds with one (or in one case two) of the new, large scaffolds.

We annotated structural and functional features of the new *L. melissa* genome using the maker pipeline (version 2.31.10)[69,70]. This pipeline uses repeat masking, protein and RNA alignment, and ab initio gene prediction to perform evidence-based gene prediction, which generates annotations that are supported by quality scores. Prior to using maker, we identified repeats in the *L. melissa* genome using repeatscout (version 1.0.5)[71]. This program identifies repeat elements, including tandem repeats and low-complexity elements, and removes them from the genome. We took this approach to avoid missing repeat elements not already in standard data bases. We provided this de novo repeat library to maker, and used this along with repbase in repeatmasker (version 4.0.7) to mask repetitive elements in the genome. Maker can use protein and RNA sequence data for genome annotation. Since we lacked protein sequences for *L. melissa*, we downloaded 28 protein sequence fasta files from 15 butterfly species (see Supplementary Table 19) from LepBase (version 4)[72]. We concatenated these fasta files to generate a protein sequence data base for maker. We used data from 24 *L. melissa* transcriptomes as additional evidence for genome annotation. We first used trimgalore (version 2.6.6, https://github.com/FelixKrueger/TrimGalore) for adapter trimming and quality filtering of paired-end RNA sequences. We then used these trimmed reads to generate a de novo transcriptome assembly with trinity (version 2.6.6)[73,74]. The assembled transcriptome was passed to maker.

We ran two rounds of the maker pipeline. We first ran maker without using any information from ab initio gene predictors (e.g., augustus), to generate de novo gene models for the *L. melissa* genome. We then ran the maker pipeline again, and used the gene models from the first run to train two gene predictors: augustus and snap. We ran snap (version 2006-07-08)[75] by using models with AED scores of 0.25 or better and a length of 50 or more amino acids. We ran augustus (version 3.3) with the insect predictions[76]. We then used both of these sets of gene predictions in the second run of maker. We then used the output from maker to obtain functional annotations of the *L. melissa* genome. We assigned putative gene functions by using blastp (version 2.3.0+) to query the maker output against the UNIPROT/SWISSPROT database[77]. We also used interproscan (version 5.32–71.0)[78] to add protein and gene ontology information to gene models. The final annotation included 11,247 putative genes, 48,765 putative coding sequences, and 8893 UTR sequences.

DNA sequencing, alignment, and genetic variant detection. We analyzed genotyping-by-sequencing (GBS) data from 835 *Lycaeides* butterflies from 23 populations: eight *L. melissa* populations (N = 238 butterflies), five *L. idas* populations (N = 176 butterflies), nine Jackson Hole *Lycaeides* populations (N = 306 butterflies), and the Dubois hybrid zone (N = 115 butterflies) (Supplementary Table 1). Samples from Yellowstone National Park and Grand Teton National Park were collected in accordance with US national park study permits YELL-05924 and GRTE-00285, respectively. The sequence data from 643 of these butterflies were previously described in a study of admixture across the *Lycaeides* species complex[28]. Data from 192 of the butterflies were generated for this study, and this includes many (but not all) of the Dubois individuals. We extracted DNA, generated genotyping-by-sequencing (GBS) libraries, and sequenced these libraries following the protocols described in ref. [28], that is, with the same protocols used for the previously published data from the 643 butterflies (a description of these methods follows). Specifically, genomic DNA was purified using Qiagen's DNeasy Blood and Tissue kit in accordance with the manufacturer's recommendations (Qiagen Inc.). Genomic DNA (6 μl per sample) was then digested with the restriction enzymes EcoRI (0.25 μL at 20,000 units per mL) and MseI (0.1 μL at 10,000 units per mL). Double-stranded DNA oligos with sequences overlapping the enzyme-cut sites and including a 8–10-bp sequence barcode and the Illumina adaptor sequences (1 μL each at 10 μM) were then ligated to the fragmented DNA with T4 DNA ligase (0.17 μL at 20,000 units per mL). We then PCR-amplified the fragment libraries with standard Illumina PCR primers (30 cycles with annealing for 30 s at 60 °C, and extension for 30 s at 72 °C followed by a final amplification with 2 min for annealing and 10 min for extension). Amplified libraries were then pooled, purified, and size-selected (300–450 bp) with a BluePippin at the Utah State University genomics core lab. The GBS libraries were sequenced on an Illumina HiSeq 2500 (100 bp, single-end reads) by the Genome Sequencing and Analysis Facility at the University of Texas (Austin, TX).

We used bwa (version 0.7.17) to align the (demultiplexed) GBS sequences from 835 individuals to the draft *L. melissa* genome by using the mem algorithm[79,80]. We ran bwa mem with a minimum seed length of 15, internal seeds of longer than 20 bp, and only output alignments with a quality score of ≥30. We then used samtools (version 1.5) to compress, sort, and index the alignments[81]. We used samtools (version 1.5) and bcftools (version 1.6) for variant calling. For variant calling, we used the recommended mapping-quality adjustment for Illumina data (−C 50), skipped alignments with mapping quality <20, skipped bases with base quality <15, and ignored insertion–deletion polymorphisms. We set the prior on SNPs to 0.001 (−P) and called SNPs when the posterior probability that the nucleotide was invariant was ≤0.01 (−p). We filtered the initial set of variants to retain only those SNPs with sequence data for at least 80% of the individuals, a mean sequence depth of two per individual, at least four reads of the alternative allele, a minimum quality score of 30, a minimum (overall) minor allele frequency of at least 0.005, and no more than 1% of the reads in the reverse orientation (this is an expectation for our GBS method). We further removed SNPs with excessive coverage (3 standard deviations above the mean) or that were tightly clustered (within 3 bp of each other), as these could reflect poor alignments (e.g., reads from multiple paralogs mapping to the same region of the genome). Finally, because we combined data from two different sequencing runs, we also removed any SNP with a difference in sequence coverage between the published and new data that was more than half the mean coverage for the two data sets combined. This left us with 39,139 SNPs for downstream analyses.

Describing patterns of genetic variation. We used the admixture model from entropy (version 1.2)[28] to obtain Bayesian estimates of genotypes and admixture proportions. This analysis was based on the full data set of 835 individuals and 39,139 SNPs. The admixture model in entropy is similar to that in structure[82], but differs by accounting for uncertainty in genotypes arising from limited sequence coverage and sequence errors, and by allowing simultaneous estimation of genotypes and admixture proportions[28]. We fit the model with $k \in \{2...5\}$ source populations. For each value of k, we ran three Markov chain Monte Carlo (MCMC) chains, each with 15,000 MCMC iterations, a burn-in of 5000 iterations and a thinning interval of 5. We used assignments from a discriminant analysis of principal components to initialize the MCMC algorithm; this speeds convergence to the posterior and avoids label switching during MCMC without affecting the posterior probability distribution[28]. We obtained genotype estimates as the posterior mean allele count for each individual and locus across chains and values of k (i.e., this integrates over uncertainty in the number of hypothetical source populations). We focused on admixture proportions for $k = 2$, as we were interested in the two nominal species and hybrids between them. We summarized patterns of population structure and admixture across the sampled populations and individuals based on these admixture proportions and a principal component analysis (PCA) of the genotypic data. We performed the PCA in R on the centered, but not standardized genotype matrix with the prcomp function.

We used the expectation–maximization (EM) algorithm implemented in estpEM (version 0.1)[57] to estimate allele frequencies for each SNP (39,139 SNPs) and population (N = 23 populations). The EM algorithm estimates allele frequencies while allowing for uncertainty in genotypes[57,83]. We used the genotype likelihoods calculated with bcftools for this analysis, and ran the algorithm with a convergence tolerance of 0.001 and allowing for 20 EM iterations. We used these

allele frequency estimates to designate ancestry-informative SNPs/markers (AIMs) as those with an allele frequency difference ≥0.3 between *L. melissa* (mean of BST, SIN, CDY, and CKV) and *L. idas* (mean of KHL, SDC, and SYC) (population IDs are defined in Supplementary Table 1).

Next, we calculated a metric of linkage disequilibrium (LD), the Pearson correlation coefficient between genotypes at pairs of loci, for all pairwise combinations of AIMs in each population where 20 or more butterflies were collected. Our goal was to ask whether LD was elevated in the Dubois hybrid zone, as predicted by theory (e.g., ref. [84]). We first polarized genotypes such that positive LD (i.e., positive Pearson correlations) coincided with an association between coupling alleles, that is, alleles more common in *L. idas* or *L. melissa*, whereas negative LD (i.e., negative Pearson correlations) coincided with associations between repulsion alleles, that is, positive associations between *L. melissa* and *L. idas* alleles. Ongoing hybridization should cause a shift toward higher positive estimates of LD, even for unlinked markers (via admixture LD). Correlations were calculated in R (version 3.5.1).

Last, we used discriminant analysis of the genetic PCs to assign butterflies from Dubois to reference taxa. We did not do this to estimate admixture proportions, but rather to determine the likely taxonomic origin of the hybridizing taxa. We first ran k-means clustering on the PC scores (PCs 1 and 2) for all sampled *Lycaeides*, except those from Dubois. We assumed three groups, allowing 50 starts and 50 iterations. This was done with the kmeans function in R. We then used the inferred clusters, which corresponded to *L. melissa*, *L. idas*, and Jackson Hole *Lycaeides* (see "Results"), as training information for a linear discriminant analysis where we classified the Dubois individuals. We assumed prior probabilities of $\frac{1}{3}$ for assignment to each cluster. This was done with the lda function in R.

Patterns of isolation-by-distance and taxon. We quantified patterns of gene flow and isolation-by-distance among the *L. idas*, *L. melissa*, and Jackson Hole *Lycaeides* populations. We especially wanted to know whether the whole system (excluding Dubois) was well described by a simple isolation-by-distance model, as might be expected for primary divergence with limited dispersal within a single taxon, or alternatively if there was evidence of restricted gene flow among these three entities, consistent with past work and with the hypothesis of secondary contact and admixture in the Jackson Hole area. We used two complementary approaches to address this question. First, we estimated relative effective migration rates among the populations based on a population allele frequency covariance matrix, which we calculated from all 39,193 SNPs. This was done using the program eems (version 0.0.0.9000)[85]. This method does not aim to estimate absolute migration rates, but rather identifies regions in space with low or high gene flow relative to a simple two-dimensional stepping-stone isolation-by-distance model[85]. Based on past work[28–30], we expected higher effective gene flow among conspecific populations, and lower effective gene flow between species and especially in the Jackson Hole admixture area. We assumed 300 demes on a triangular grid, and fit the model using MCMC with three chains, 6 million sampling iterations, 3 million burn-in iterations, and a thinning interval of 10,000.

Next, we fit Bayesian linear mixed models to assess the relative contributions of geographic distance and differences in nominal taxon (*L. idas*, *L. melissa*, or Jackson Hole *Lycaeides*) in explaining patterns of genetic differentiation among populations (as described in ref. [28]). This Bayesian regression analysis is based on the mixed model framework proposed by Clarke et al.[86] to account for the correlated error structure inherent in pairwise observations such as genetic distances. We fit a model for logit-transformed F_{ST} (defined as $\frac{1/L\Sigma_L(H_T - H_S)}{1/L\Sigma_L H_T}$, where H_S and H_T are the expected heterozygosities for the individual and combined population pairs) as a function of log geographic distance (great circle distance) and taxon distance (defined as 0 for populations from the same nominal taxon and 1 for populations from different nominal taxa). We fit the Bayesian model using MCMC via the rjags (version 4.8) interface with JAGS (version 4.3.0). We placed minimally informative priors on the covariates, normal ($\mu = 0$, $\tau = 0.001$), and on the population random effects and residual errors, both gamma (1, 0.01). We compared the full models with models with only geographic or taxon distance using deviance information criterion (DIC). We ran three chains for each model, each with 10,000 iterations, a 2000-iteration burn-in, and a thinning interval of 5. A model, including geographic distance and the same versus different taxon, was preferred (DIC = 92.82) relative to the one with only taxon (DIC = 98.67) or only geographic distance (DIC = 360.9).

Analysis of population ancestry in Jackson Hole. We estimated ancestry segment frequencies across the genome for each of the nine Jackson Hole ancient hybrid populations using the correlated beta-process model implemented in popanc (ver. 0.1)[15]. This method is similar to a hidden Markov model and accounts for the expected autocorrelation in ancestry along chromosomes, but allows ancestry frequencies to vary along the genome. It is particularly well suited for cases where genome stabilization has begun but is not yet complete (i.e., where populations consist of individuals with similar admixture proportions but where ancestry frequencies vary across the genome)[15]. We ran popanc for each of the Jackson Hole populations to estimate the frequency of *L. idas*-derived alleles along the genome. We used a window size of three SNPs, and focused on the 1164 AIMs. We set SYC and KHL as representative of the *L. idas* parental species, and BST,

SIN, CDY, and CKV as the putative *L. melissa* parents. Maximum likelihood allele frequency estimates from estpEM were used as input for the program/analysis. We ran the MCMC analysis for each population with a 10,000-iteration chain, a 5000-iteration burn-in, and thinning interval of 10. We based our inferences on point estimates (posterior mean) of *L. idas* ancestry frequencies for individual populations or on averages of these estimates across all nine populations. We repeated this analysis with only males and with AIMs defined as those SNPs with an allele frequency difference of ≥0.2 between *L. melissa* and *L. idas* to assess the robustness of our results. Randomization tests were used to ask whether and to what extent AIMs with the highest *L. idas* or *L. melissa* ancestry frequencies (top 10% in each case) were overrepresented on the Z chromosome or in or near certain structural features of the genome, such as genes, coding sequences, transposable elements, and annotated protein sequences or motifs. Null expectations were derived from 1000 permutations of ancestry frequency estimates across the 1164 AIMs. Last, we asked whether the largest chromosomes harbored more *L. idas* ancestry than smaller chromosomes (as noted in the Results, Jackson Hole *Lycaeides* inherited more of their genomes from *L. idas* than *L. melissa*). Such a pattern is expected when selection acts on many loci and against alleles from the minor parent (the one that contributes less to overall ancestry, in this case *L. melissa*) and when recombination (per bp) is lower on larger chromosomes (as has been seen in Lepidoptera), as neutral alleles from the minor parent then have less opportunity to recombine away from deleterious alleles from the minor parent when they occur on larger chromosomes[25].

Estimating cline parameters in the Dubois hybrid zone. We fit Bayesian genomic clines for each of the 1164 AIMs using bgc (ver. 1.04b[33]) to quantify genome-wide variability in introgression between *L. melissa* and Jackson Hole *Lycaeides* in the Dubois hybrid zone. This method estimates clines in ancestry for individual genetic loci (e.g., SNPs) along a genome-average admixture gradient[14,34]. As such, it can be applied in cases where hybrid zones are confined to a single geographic locality, such as the Dubois population. Unlike the method used for the ancient Jackson Hole hybrids, the genomic cline method performs best when a wide range of hybrids with different genome compositions exist, as is the case for the Dubois hybrid zone. Deviations between genome-average introgression and introgression for each locus are measured with two cline parameters, α and β. Cline parameter α denotes an increase (for positive values of α) or decrease (for negative values of α) in the probability of ancestry from reference (parental) species 1 relative to null expectations from an individual's hybrid index. Genomic cline parameter β describes an increase (positive values) or decrease (negative values) in the rate of transition from parental species 0 ancestry to parental species 1 ancestry along the genome-wide admixture gradient. When placed in a geographic context, α is equivalent to twice the shift in cline center, and β measures the decrease (or increase when $\beta < 0$) in cline width relative to the average[14,84]. Cline parameters can be affected by genetic drift and selection in hybrids[14,35]. However, in the absence of major geographic barriers to gene flow, high positive values of β (i.e., the equivalent of narrow clines in a geographic context) are most readily explained by selection.

We fit the bgc genomic cline model for the 115 *Lycaeides* butterflies from the Dubois hybrid zone. We used *L. melissa* (LAN, SIN, and CKV, $N = 131$ butterflies) and Jackson Hole *Lycaeides* (BLD and FRC, $N = 94$ butterflies) as reference or parental species for the analysis. These specific populations were chosen as the source populations because they were nearest to the Dubois hybrid zone. Using these populations as references, allele frequencies for the AIMs differed sufficiently between source populations to obtain meaningful estimates of ancestry (e.g., mean = 0.28, 78% with allele frequency differences >0.1, see also simulations in ref. [35]). Genotype likelihoods from bcftools were used as input for the analysis (the model fit incorporates uncertainty in genotype as captured by the genotype likelihoods). We ran the analysis using only the 1164 AIMs. We fit the model using MCMC, with five chains each with 25,000 iterations, a 5000-iteration burn-in, and a thinning interval of 5. We inspected the MCMC output to assess convergence of chains to the stationary distribution and combined the output of the five chains. We repeated this analysis with only males and with AIMs defined as those SNPs with an allele frequency difference of ≥0.2 between *L. melissa* and *L. idas* (2126 SNPs) to assess the robustness of our results.

We defined credible deviations from null expectations given genome-wide admixture as cases where the 95% credible intervals (specifically the equal-tail probability intervals) for α or β for a given locus (AIM) excluded 0. We focused specifically on cases of credible directional ($\alpha \neq 0$) or restricted ($\beta > 0$) introgression relative to the genome-wide average. We used randomization tests to ask whether (and to what extent) such loci were over (or under-) represented on the Z chromosome or in or near (within 1000 bp of) certain annotated structural features of the genome, such as genes, coding sequences, transposable elements, and annotated protein sequences or motifs. Null expectations were derived from 1000 permutations of cline parameter estimates across the 1164 AIMs. We also tested for correlations between chromosome size and the proportion of AIMs showing evidence of restricted ($\beta > 0$) or directional ($\alpha > 0$ or $\alpha < 0$) introgression.

Quantifying consistency in outcomes of hybridization. We next tested for excess overlap between AIMs showing the most extreme ancestry frequencies in the ancient Jackson Hole hybrids and those with the greatest deviations from genome-wide

average introgression in the contemporary Dubois hybrid zone. We focused primarily on the following six comparisons: (i) excess introgression of Jackson Hole alleles in Dubois ($\alpha > 0$) and high *L. idas* ancestry in Jackson Hole, (ii) excess introgression of Jackson Hole alleles in Dubois ($\alpha > 0$) and high *L. melissa* ancestry in Jackson Hole, (iii) excess introgression of *L. melissa* alleles in Dubois ($\alpha < 0$) and high *L. idas* ancestry in Jackson Hole, (iv) excess introgression of *L. melissa* alleles in Dubois ($\alpha < 0$) and high *L. melissa* ancestry in Jackson Hole, (v) restricted introgression in Dubois ($\beta > 0$) and high *L. idas* ancestry in Jackson Hole, and (vi) restricted introgression in Dubois ($\beta > 0$) and high *L. melissa* ancestry in Jackson Hole. We were especially interested in the last two comparisons (v and vi), as the restricted introgression AIMs constitute our best candidates for regions of the genome associated with reproductive isolation. We initially focused on the top 10% of AIMs in each of the categories (e.g., the 116 out of 1164 AIMs with the highest point estimates of α from bgc). We conducted randomization tests by permuting parameters across AIMs to test whether and to what extent AIMs in the top 10% for each comparison coincided more for each comparison than expected by chance. We conducted 10,000 permutations to generate null expectations. The null distribution was used to calculate a P-value and x-fold enrichment for each comparison. As an example, an x-fold enrichment of 2.0 would indicate that twice as many AIMs exhibited a pair of patterns (e.g., restricted introgression in Dubois and high *L. idas* ancestry frequencies in Jackson Hole) as expected by chance (based on the mean of the null). We repeated these analyses considering the top 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, and 1% of AIMs in each category (i.e., across more and more extreme quantiles), and with only autosomal or only Z-linked AIMs. We also repeated this analysis with only males and with AIMs defined as those SNPs with an allele frequency difference of ≥0.2 between *L. melissa* and *L. idas* to further assess the robustness of our results. As a final assessment of robustness, we considered three additional comparisons: (vii) excess introgression of Jackson Hole alleles in Dubois ($\alpha > 0$) and high *L. idas* or *L. melissa* ancestry in Jackson Hole (i.e., genomic regions nearest fixation for either ancestry type, hereafter extreme ancestry), (viii) excess introgression of *L. melissa* alleles in Dubois ($\alpha < 0$) and extreme ancestry in Jackson Hole, and (ix) restricted introgression in Dubois ($\beta > 0$) and extreme ancestry in Jackson Hole. We include these comparisons as the direction of selection in Jackson Hole (for or against *L. idas* or *L. melissa* ancestry) might not predict the direction of selection in Dubois, but still the same subsets of loci could be affected by selection in both cases.

Finally, we extracted IPR and GO terms and descriptions generated by maker for the set of AIMs in the top 10% for each comparison where we had evidence of significantly greater overlap between categories than expected by chance. We filtered the IPR terms based on their hierarchical classifications (i.e., superfamily, family, domain, or repeat). We retained unique terms and dropped terms that are at lower or overlapping levels in the Interproscan database[78]. For example, if a SNP was annotated for a superfamily IPR term and multiple domains within the superfamily, we retained only the superfamily term.

Whole-genome phylogenomic analyses. We next asked whether patterns of introgression in the contemporary, Dubois hybrid zone were also predictive of patterns of ancestry in two additional, ancient hybrid lineages from the Sierra Nevada and Warner mountains of western North America[6,28,32]. To do this, we generated whole-genome sequences from *L. anna* (from Yuba Gap, CA), *L. melissa* (from Bonneville Shoreline, UT), *L. idas* (from Trout Lake, WY), and the Sierra Nevada (from Carson Pass, NV), and Warner mountain (from Buck Mountain, CA) lineages. We extracted DNA from one female (ZW) butterfly per population using Qiagen's MagAttract HMW DNA extraction kit (Qiagen, Inc.) following the manufacturer's suggested protocol. We then outsourced library construction and sequencing to Macrogen Inc. (Seoul, South Korea). One standard paired-end shotgun library (180-bp insert) was constructed for each butterfly/lineage using a TruSeq library preparation kit (Illumina, Inc.). Each library was sequenced on its own lane on a HiSeq 2000 with 2×100-bp paired reads. We obtained >20 million high-quality (≥Q30) reads for each library (~8X coverage of each genome).

We ran bwa mem with a minimum seed length of 15, internal seeds of longer than 20 bp, and only output alignments with a quality score of ≥30. We then used samtools (version 1.5) to compress, sort, and index the alignments, and to remove PCR duplicates[81]. We used GATK's HaplotypeCaller (version 3.5) to call SNPs across the five genomes[87]. We excluded bases with mapping base-quality scores less than 30, assumed a prior heterozygosity of 0.001, applied the aggressive PCR indel model, and only called variants with a minimum confidence threshold of 50. We used the intermediate g.vcf file approach followed by joint variant calling with the GenotypeGVCFs command when calling variants. We then filtered the initial set of variants to include only SNPs with an average minimum coverage of 6×, maximum absolute values of 2.5 for the base-quality rank-sum test, the mapping-quality rank-sum test and the read position rank-sum test, a minimum ratio of variant confidence to non-reference read depth of 2.5, and a minimum mapping quality of 30. We also excluded all indels and SNPs with more than two alleles, and all SNPs on smaller scaffolds not assigned to linkage groups. Finally, SNPs with exceptionally high coverage (>450×) or that were clustered together (within 5 bps of each other) were dropped. This left us with 2,054,096 SNPs for phylogenomic analyses.

We first estimated a genome consensus or "species" tree from a concatenated alignment of the autosomal SNPs (i.e., this analysis did not include SNPs on the

Z chromosome). We estimated the tree with RAxML (version 8.2.9)[88] under the GTR model with no rate heterogeneity. In total, 100 bootstrap replicates were generated and analyzed to assess confidence in the bifurcations in the estimated phylogeny. Next, to examine variation in (unrooted) tree topologies across the genome, we split the alignment into nonoverlapping 1000 SNP windows. We estimated phylogenies for each window using RAxML as described above. We then used the R packages ape (version 5.2)[89] and phytools (version 0.6.60)[90] to identify trees (across windows) with the same topology and to quantify the proportion/ number of trees on each linkage group with each topology. We used randomization tests to determine whether autosomes, the Z chromosome, or the 49 candidate barrier loci (i.e., regions containing the 49 AIMs with restricted introgression in Dubois and high *L. idas* ancestry frequencies in Jackson Hole) exhibited a significant excess or deficit of a given topology. This was done in R as well. We used 1000 randomizations (permutations) of tree topologies across 1000 SNP windows (and thus linkage groups) for each analysis. Finally, we tested for correlations between the frequency of specific topologies (i.e., an "introgression" tree and tree uniting the ancient hybrid lineages) and chromosome size.

Variation in the frequency of different topologies across the genome, and particularly in the prevalence of the species-tree topology, can arise from introgression or incomplete lineage sorting with differences in effective population sizes across chromosomes (e.g., ref. [25]). Thus, as an explicit test for reduced introgression on the Z chromosome, we calculated f_d admixture proportions for each linkage group[48]. We assumed that the species topology for this calculation was ((*L. idas* and *L. melissa*) (H, *L. anna*), where H denotes either the Sierra Nevada or Warner mountain lineage. f_d was estimated following ref. [48] using our own script written in R. We then tested for correlations between admixture at the chromosome level (f_d) and chromosome size. Finally, we asked whether the set of 1000 SNP windows containing the 49 candidate barrier loci showed less evidence of admixture than expected by chance. For this, we calculated f_d for the set of 49, 1000 SNP windows and compared this with a null distribution generated by repeatedly sampling 49, 1000-bp windows at random, and computed f_d for the sampled set of windows. This was done in R.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

DNA-sequence data that support the findings of this study have been deposited in the NCBI SRA with accession codes PRJNA577236 and PRJNA432816. The *L. melissa* genome assembly, annotation, RNA-sequence data used for the annotation, and SNP variant file that also support the findings of this study have been deposited in Dryad[91]. Publicly available databases that support the findings of this study are UNIPROT/ SWISSPROT (http://www.uniprot.org) and LepBase version 4.0 (http://lepbase.org/). The source data underlying Figs. 2–6 are provided as a Source Data file.

Code availability

Scripts and computer code used in the analyses of these data are available on Dryad (https://doi.org/10.5061/dryad.76hdr7ssw)[92] and from GitHub.

References

1. Arnold, M. L. *Natural Hybridization and Evolution* (Oxford University Press, 1997).
2. Mallet, J. Hybridization as an invasion of the genome. *Trends Ecol. Evolution* **20**, 229–237 (2005).
3. Prüfer, K. et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43 (2014).
4. De Manuel, M. et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* **354**, 477–481 (2016).
5. Taylor, S. A. & Larson, E. L. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nat. Ecol. Evolution* **3**, 170 (2019).
6. Gompert, Z., Fordyce, J. A., Forister, M. L., Shapiro, A. M. & Nice, C. C. Homoploid hybrid speciation in an extreme habitat. *Science* **314**, 1923–1925 (2006).
7. Buerkle, C. A. & Rieseberg, L. H. The rate of genome stabilization in homoploid hybrid species. *Evolution* **62**, 266–275 (2008).
8. Sankararaman, S. et al. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354 (2014).
9. Schumer, M., Cui, R., Powell, D. L., Rosenthal, G. G. & Andolfatto, P. Ancient hybridization and genomic stabilization in a swordtail fish. *Mol. Ecol.* **25**, 2661–2679 (2016).
10. Elgvin, T. O. et al. The genomic mosaicism of hybrid speciation. *Sci. Adv.* **3**, e1602996 (2017).
11. Lamichhaney, S. et al. Rapid hybrid speciation in Darwinas finches. *Science* **359**, 224–228 (2018).
12. Barton, N. H. & Hewitt, G. M. Analysis of hybrid zones. *Annu. Rev. Ecol. Syst.* **16**, 113–148 (1985).
13. Abbott, R. J., Barton, N. H. & Good, J. M. Genomics of hybridization and its evolutionary consequences. *Mol. Ecol.* **25**, 2325–2332 (2016).
14. Gompert, Z., Mandeville, E. G. & Buerkle, C. A. Analysis of population genomic data from hybrid zones. *Annu. Rev. Ecol., Evolution, Syst.* **48**, 207–229 (2017).
15. Gompert, Z. A continuous correlated beta process model for genetic ancestry in admixed populations. *PLoS ONE* **11**, e0151047 (2016).
16. Schumer, M. et al. Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science* **360**, 656–660 (2018).
17. Barton, N. & Bengtsson, B. O. The barrier to genetic exchange between hybridising populations. *Heredity* **57**, 357 (1986).
18. Rieseberg, L. H., Whitton, J. & Gardner, K. Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics* **152**, 713–727 (1999).
19. Payseur, B. A., Krenz, J. G. & Nachman, M. W. Differential patterns of introgression across the x chromosome in a hybrid zone between two species of house mice. *Evolution* **58**, 2064–2078 (2004).
20. Harrison, R. G. & Larson, E. L. Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. *Mol. Ecol.* **25**, 2454–2466 (2016).
21. Payseur, B. A. & Rieseberg, L. H. A genomic perspective on hybridization and speciation. *Mol. Ecol.* **25**, 2337–2360 (2016).
22. Rieseberg, L. H. et al. Major ecological transitions in wild sunflowers facilitated by hybridization. *Science* **301**, 1211–1216 (2003).
23. Brandvain, Y., Kenney, A. M., Flagel, L., Coop, G. & Sweigart, A. L. Speciation and introgression between Mimulus nasutus and Mimulus guttatus. *PLoS Genet.* **10**, e1004410 (2014).
24. Juric, I., Aeschbacher, S. & Coop, G. The strength of selection against Neanderthal introgression. *PLoS Genet.* **12**, e1006340 (2016).
25. Martin, S. H., Davey, J. W., Salazar, C. & Jiggins, C. D. Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biol.* **17**, e2006288 (2019).
26. Mandeville, E. G. et al. Inconsistent reproductive isolation revealed by interactions between Catostomus fish species. *Evolution Lett.* **1**, 255–268 (2017).
27. Forister, M. L., Gompert, Z., Fordyce, J. A. & Nice, C. C. After 60 years, an answer to the question: what is the Karner blue butterfly? *Biol. Lett.* **7**, 399–402 (2011).
28. Gompert, Z. et al. Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Mol. Ecol.* **23**, 4555–4573 (2014).
29. Gompert, Z., Lucas, L. K., Fordyce, J. A., Forister, M. L. & Nice, C. C. Secondary contact between Lycaeides idas and L. melissa in the Rocky Mountains: extensive admixture and a patchy hybrid zone. *Mol. Ecol.* **19**, 3171–3192 (2010).
30. Gompert, Z. et al. Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution* **66**, 2167–2181 (2012).
31. Gompert, Z. et al. Geographically multifarious phenotypic divergence during speciation. *Ecol. Evolution* **3**, 595–613 (2013).
32. Nice, C. C. et al. Hybrid speciation and independent evolution in lineages of alpine butterflies. *Evolution* **67**, 1055–1068 (2013).
33. Gompert, Z. & Buerkle, C. bgc: Software for Bayesian estimation of genomic clines. *Mol. Ecol. Resour.* **12**, 1168–1176 (2012).
34. Gompert, Z. & Buerkle, C. A. Bayesian estimation of genomic clines. *Mol. Ecol.* **20**, 2111–2127 (2011).
35. Gompert, Z., Parchman, T. L. & Buerkle, C. A. Genomics of isolation in hybrids. *Philos. Trans. R. Soc. B: Biol. Sci.* **367**, 439–450 (2012).
36. Stanic, B. et al. Cold hardiness in Ostrinia nubilalis (Lepidoptera: Pyralidae): glycerol content, hexose monophosphate shunt activity, and antioxidative defense system. *Eur. J. Entomol.* **101**, 459–466 (2004).
37. Fan, L., Lin, J., Zhong, Y. & Liu, J. Shotgun proteomic analysis on the diapause and non-diapause eggs of domesticated silkworm Bombyx mori. *PLoS One* **8**, e60386 (2013).
38. Adrion, J. R., Hahn, M. W. & Cooper, B. S. Revisiting classic clines in Drosophila melanogaster in the age of genomics. *Trends Genet.* **31**, 434–444 (2015).
39. Hagen, R. & Scriber, J. Sex-linked diapause, color, and allozyme loci in Papilio glaucus: linkage analysis and significance in a hybrid zone. *J. Heredity* **80**, 179–185 (1989).
40. Scriber, J. M., Giebink, B. L. & Snider, D. Reciprocal latitudinal clines in oviposition behavior of Papilio glaucus and P. canadensis across the Great

Lakes hybrid zone: possible sex-linkage of oviposition preferences. *Oecologia* **87**, 360–368 (1991).

41. Cong, Q., Borek, D., Otwinowski, Z. & Grishin, N. V. Tiger swallowtail genome reveals mechanisms for speciation and caterpillar chemical defense. *Cell Rep.* **10**, 910–919 (2015).

42. Mandrioli, M., Monti, M. & Tedeschi, R. Presence and conservation of the immunoglobulin superfamily in insects: current perspective and future challenges. *Invertebr. Survival J.* **12**, 188–194 (2015).

43. Teeter, K. C. et al. Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Res.* **18**, 67–76 (2008).

44. Ozaki, K., Utoguchi, A., Yamada, A. & Yoshikawa, H. Identification and genomic structure of chemosensory proteins (CSP) and odorant binding proteins (OBP) genes expressed in foreleg tarsi of the swallowtail butterfly *Papilio xuthus. Insect Biochem. Mol. Biol.* **38**, 969–976 (2008).

45. Briscoe, A. D. et al. Female behaviour drives expression and evolution of gustatory receptors in butterflies. *PLoS Genet.* **9**, e1003620 (2013).

46. Martin, A. et al. Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand. *Proc. Natl Acad. Sci. USA* **109**, 12632–12637 (2012).

47. Gompert, Z., Fordyce, J. A., Forister, M. L. & Nice, C. C. Recent colonization and radiation of North American Lycaeides (Plebejus) inferred from mtDNA. *Mol. Phylogenet. Evol.* **48**, 481–490 (2008).

48. Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evolution* **32**, 244–257 (2014).

49. Gould, N. E.-S. J. & Eldredge, N. Punctuated equilibria: an alternative to phyletic gradualism.(eds. Ayala, F. J. & Avise, J. C.) in *Essential Readings in Evolutionary Biology* 82–115 (Johns Hopkins University Press, 1972).

50. Reznick, D. N. & Ricklefs, R. E. Darwinas bridge between microevolution and macroevolution. *Nature* **457**, 837 (2009).

51. Wagner, C. E., Harmon, L. J. & Seehausen, O. Ecological opportunity and sexual selection together predict adaptive radiation. *Nature* **487**, 366 (2012).

52. Rieseberg, L. H., Sinervo, B., Linder, C. R., Ungerer, M. C. & Arias, D. M. Role of gene interactions in hybrid speciation: evidence from ancient and experimental hybrids. *Science* **272**, 741–745 (1996).

53. DiFazio, S. P., Slavov, G. T., Joshi, C. P. et al. *Populus*: a premier pioneer system for plant genomics. (eds. Joshi, C. P., DiFazio, S. P. & Kole, C.) in *Genetics, Genomics and Breeding of Poplar.* 1–28 (Science Publishers, Enfield, NH, 2011).

54. Baute, G. J., Owens, G. L., Bock, D. G. & Rieseberg, L. H. Genome-wide genotyping-by-sequencing data provide a high-resolution view of wild Helianthus diversity, genetic structure, and interspecies gene flow. *Am. J. Bot.* **103**, 2170–2177 (2016).

55. Chhatre, V. E., Evans, L. M., Di Fazio, S. P. & Keller, S. R. Adaptive introgression and maintenance of a trispecies hybrid complex in range-edge populations of Populus. *Mol. Ecol.* **27**, 4820–4838 (2018).

56. Endler, J. A. *Natural Selection in the Wild.* 21 (Princeton University Press, 1986).

57. Soria-Carrasco, V. et al. Stick insect genomes reveal natural selectionas role in parallel speciation. *Science* **344**, 738–742 (2014).

58. Turelli, M., Barton, N. H. & Coyne, J. A. Theory and speciation. *Trends Ecol. evolution* **16**, 330–343 (2001).

59. Coyne, J. A. & Allen, O. H. *Speciation* (Sinauer Associates, Inc., 2004).

60. Masly, J. P. & Presgraves, D. C. High-resolution genome-wide dissection of the two rules of speciation in Drosophila. *PLoS Biol.* **5**, e243 (2007).

61. Muirhead, C. A. & Presgraves, D. C. Hybrid incompatibilities, local adaptation, and the genomic distribution of natural introgression between species. *Am. Naturalist* **187**, 249–261 (2016).

62. Counterman, B. A. et al. Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in Heliconius erato. *PLoS Genet.* **6**, e1000796 (2010).

63. Martin, A. & Orgogozo, V. The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution* **67**, 1235–1250 (2013).

64. Gompert, Z. et al. The evolution of novel host use is unlikely to be constrained by trade-offs or a lack of genetic variation. *Mol. Ecol.* **24**, 2777–2793 (2015).

65. Chaturvedi, S. et al. The predictability of genomic changes underlying a recent host shift in Melissa blue butterflies. *Mol. Ecol.* **27**, 2651–2666 (2018).

66. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598 (2015).

67. Putnam, N. H. et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).

68. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478–2483 (2002).

69. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**, 329 (2012).

70. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome annotation and curation using maker and maker-p. *Curr. Protoc. Bioinforma.* **48**, 4–11 (2014).

71. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).

72. Challis, R. J., Kumar, S., Dasmahapatra, K. K. K., Jiggins, C. D. & Blaxter, M. Lepbase: the lepidopteran genome database. Preprint at https://doi.org/10.1101/056994 (2016).

73. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644 (2011).

74. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494 (2013).

75. Korf, I. Gene finding in novel genomes. *BMC Bioinforma.* **5**, 59 (2004).

76. Stanke, M. & Morgenstern, B. Augustus: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).

77. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2018).

78. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

79. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

80. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).

81. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

82. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).

83. Li, H. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

84. Szymura, J. M. & Barton, N. H. Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegata*, near Cracow in southern Poland. *Evolution* **40**, 1141–1159 (1986).

85. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94 (2016).

86. Clarke, R. T., Rothery, P. & Raybould, A. F. Confidence limits for regression relationships between distance matrices: estimating gene flow with distance. *J. Agric., Biol., Environ. Stat.* **7**, 361 (2002).

87. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

88. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

89. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2018).

90. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evolution* **3**, 217–223 (2012).

91. Gompert, Z. et al. Data from: recent hybrids recapitulate ancient hybrid outcomes. *Dryad, Dataset* https://doi.org/10.5061/dryad.j3tx95x9d (2020).

92. Gompert, Z. et al. Code from: recent hybrids recapitulate ancient hybrid outcomes. *Dryad, Dataset* https://doi.org/10.5061/dryad.76hdr7ssw (2020).

Acknowledgements

Author contributions

S.C. generated and analyzed the GBS data, and wrote the paper. L.K.L. collected samples, generated some of the GBS data, and edited the paper. C.A.B., J.A.F., and M.L.F. collected samples and edited the paper. C.C.N. collected samples, generated some of the GBS data, generated the whole-genome sequence data, and edited the paper. Z.G. collected samples, generated and analyzed the GBS and the whole-genome sequence data, and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-15641-x.

Correspondence and requests for materials should be addressed to Z.G.

Peer review information *Nature Communications* thanks Takeshi Kawakami, Simon Martin and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.