MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 4555–4573 doi: 10.1111/mec.12811

Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants

ZACHARIAH GOMPERT,* LAUREN K. LUCAS,*† C. ALEX BUERKLE,‡ MATTHEW L. FORISTER,§ JAMES A. FORDYCE¶ and CHRIS C. NICE†

*Department of Biology, Utah State University, Logan, UT 84322, USA, †Department of Biology, Texas State University, San Marcos, TX 78666, USA, ‡Department of Botany and Program in Ecology, University of Wyoming, Laramie, WY 82071, USA, §Department of Biology, University of Nevada, Reno, NV 89557, USA, ¶Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN 37996, USA

Abstract

Detailed information about the geographic distribution of genetic and genomic variation is necessary to better understand the organization and structure of biological diversity. In particular, spatial isolation within species and hybridization between them can blur species boundaries and create evolutionary relationships that are inconsistent with a strictly bifurcating tree model. Here, we analyse genome-wide DNA sequence and genetic ancestry variation in *Lycaeides* butterflies to quantify the effects of admixture and spatial isolation on how biological diversity is organized in this group. We document geographically widespread and pervasive historical admixture, with more restricted recent hybridization. This includes evidence supporting previously known and unknown instances of admixture. The genome composition of admixed individuals varies much more among than within populations, and tree- and genetic ancestry-based analyses indicate that multiple distinct admixed lineages or populations exist. We find that most genetic variants in *Lycaeides* are rare (minor allele frequency <0.5%). Because the spatial and taxonomic distributions of alleles reflect demographic and selective processes since mutation, rare alleles, which are presumably younger than common alleles, were spatially and taxonomically restricted compared with common variants. Thus, we show patterns of genetic variation in this group are multifaceted, and we argue that this complexity challenges simplistic notions concerning the organization of biological diversity into discrete, easily delineated and hierarchically structured entities.

Keywords: ancestry, differentiation, genomics, hybridization

Received 8 November 2013; revision received 27 April 2014; accepted 29 April 2014

Introduction

Detailed information regarding the distribution of genetic and genomic variation within and among populations and species is needed to better understand the nature of species boundaries and the organization of biological diversity. Along these lines, limited dispersal and spatially variable selection create population-genetic structure within species (Wright 1943; Ehrlich & Raven 1969; Lee & Mitchell-Olds 2011; Wang *et al.* 2013), and admixture and introgression result in gene flow between species (Endler 1977; Barton 2001; Seehausen *et al.* 2008). Genetic exchange can erode accumulated genetic differences between species (Levin *et al.* 1996; Taylor *et al.* 2006; Fitzpatrick & Shaffer 2007; Vonlanthen *et al.* 2012), or be an important source of favourable genetic variants or novel combinations of alleles, and thereby facilitate adaptation or drive speciation (e.g. Rieseberg *et al.* 2003; Grant *et al.* 2004; Gompert *et al.* 2006; Mavárez *et al.*

Correspondence: Zachariah Gompert;
E-mail: zach.gompert@usu.edu

© 2014 John Wiley & Sons Ltd

2006; Whitney *et al.* 2006; Jones *et al.* 2012). These processes can blur species boundaries and create evolutionary relationships that vary across the genome or that are not well-represented by a strictly bifurcating evolutionary model (Harrison & Rand 1989; Wu 2001; Mallet *et al.* 2007; Mallet 2008; Dasmahapatra *et al.* 2012).

Many studies have documented population-genetic structure or introgression based on tens or hundreds of genetic markers, but we have only recently begun to investigate how these processes shape genetic variation at the genome scale (e.g. Dasmahapatra *et al.* 2012; Ellegren *et al.* 2012; Gompert *et al.* 2012; Jones *et al.* 2012; Nosil *et al.* 2012; Parchman *et al.* 2013). Genome-scale data are needed for accurate and precise inferences of population structure or historical demographic processes (e.g. Edwards & Beerli 2000; Luikart *et al.* 2003; Gompert *et al.* 2012) and to determine whether and to what extent the genetic consequences of specific historical events or evolutionary processes vary across the genome or among different classes of genetic variants (e.g. Voight *et al.* 2006; Begun *et al.* 2007; Coop *et al.* 2009; Buerkle *et al.* 2011). For example, recent large-scale resequencing studies in humans have identified an abundance of rare variants, including many rare functional variants (Gravel *et al.* 2011; Nelson *et al.* 2012). In particular, Gravel *et al.* (2011) documented segregating variation at one in every 17 nucleotide positions in human populations, with a minor allele frequency (MAF) at the majority of these sites <0.5%. Compared with common variants, rare variants in humans are geographically localized and more likely to be deleterious (Li *et al.* 2010; Mathieson & McVean 2012; Nelson *et al.* 2012). Hence, rare and common genetic variants might reflect different population boundaries and demographic or evolutionary histories, but few studies have addressed this question in non-human populations.

In this study, we analyse the geographic distribution of genomic variation in a species complex of *Lycaeides* butterflies to advance our understanding of the organization and structure of biological diversity. We previously documented patterns of genetic and morphological variation in *Lycaeides* that were consistent with the hypothesis that admixture and introgression have been common and potentially important drivers of evolution in this group of butterflies (Gompert, *et al.*, 2006, 2008; Lucas *et al.* 2008; Gompert *et al.* 2012; Nice *et al.* 2013). But each of these studies involved limited genetic information (Gompert, *et al.*, 2006, 2008), limited geographic sampling (Gompert, *et al.*, 2006, 2012), or lacked fine spatial or individual-level resolution (Gompert *et al.* 2010a; Nice *et al.* 2013). Thus, we know relatively little about geographic variation in genome composition within and among admixed lineages or their likely parental species and by extension whether these lineages are cohesive evolutionary entities. Such knowledge is critical for understanding the organization and structure of biological diversity.

Here, we overcome the limitations of our previous work and generate and analyse individual-based, genome-wide DNA sequence data from 1536 *Lycaeides* butterflies from 66 locations, including multiple localities within the range of each putative parental and admixed lineage. We use these data and analyses to answer three focal questions: (i) How well do the spatial distributions of rare (MAF between 0.1% and 0.5%), low-frequency (MAF between 0.5% and 5%) and common genetic variants (MAF >5%) correspond with taxonomic boundaries or geographic distances among populations?, (ii) Is the evolutionary history of *Lycaeides* butterflies well-described by a strictly bifurcating tree or do genome-wide sequence data support the hypothesis that putative admixed lineages are indeed of hybrid origin? and (iii) How is genetic ancestry organized within admixed individuals, and how much does the genome composition of admixed individuals vary within and among populations and lineages? Answering this last question allows us to distinguish between contemporary and historical admixture. In addition, by combining extensive individual, population and genomic sampling, we are able to describe patterns of genetic structure at a scale and level of detail that has rarely been accomplished outside of human population genetics.

Methods

Study system

Lycaeides is a holartic genus of small blue butterflies (Nabokov, 1943, 1949). These butterflies are host plant specialists on legumes, with most populations feeding exclusively on one or a few plant species (Scott 1986; Forister *et al.* 2013; Gompert *et al.* 2013b). *Lycaeides* butterflies have a patchy spatial distribution that is determined by the availability of suitable host plants (Gompert *et al.* 2010b; Forister *et al.* 2011b), with limited dispersal among localities (i.e. dispersal beyond 500 m is rare; U.S. Fish and Wildlife Service 2003). In this study, we focused on three of four nominal species in North America, *Lycaeides anna*, *L. idas* and *L. melissa* (we excluded the endangered Karner blue butterfly, *L. samuelis*; Forister *et al.* 2011a, Table 1, Fig. 1a). *Lycaeides anna* is found from California to British Columbia on the western slopes of the Sierra Nevada and in the Cascade mountains (Nabokov 1943; Guppy & Shepard 2001). A phenotypically distinct subspecies, *L. anna ricei*, is recognized in the northern portion of this range.

© 2014 John Wiley & Sons Ltd

Table 1 Locality information and sample sizes. Here and in other figures and tables, we use shortened taxon names: *L. anna* = anna, *L. anna ricei* = ricei, *L. idas* = idas, *L. melissa* = melissa, Alpine *Lycaeides* = alpine, Warner *Lycaeides* = warner and Jackson *Lycaeides* = jackson

Locality no.	Locality	Taxon	Subgroup	Longitude (°W)	Latitude (°N)	No. of individuals
1	Donner Pass, CA	anna	n/a	120.35	39.31	17
2	Fall Creek, CA	anna	n/a	120.67	39.38	20
3	Leek Springs, CA	anna	n/a	120.24	38.63	20
4	Yuba Gap, CA	anna	n/a	120.60	39.32	20
5	Castle Peak, CA	anna	n/a	120.35	39.37	16
6	Shovel Creek, CA	ricei	n/a	122.16	41.88	21
7	Big Lake, OR	ricei	n/a	121.87	44.38	20
8	Marble Mts., CA	ricei	n/a	122.75	41.83	12
9	Soda Mt. Road, OR	ricei	n/a	122.48	42.12	19
10	Cave Lake, CA	ricei	n/a	120.21	41.98	23
11	Strawberry Mts., OR	idas	n/a	118.64	44.34	20
12	King's Hill, MT	idas	n/a	110.70	46.84	18
13	Soldier Creek, MT	idas	idas-north	114.61	47.21	19
14	Siyeh Creek, MT	idas	idas-north	113.67	48.70	20
15	Bunsen Peak, WY	idas	idas-ynp	110.72	44.93	20
16	Garnet Peak, MT	idas	idas-ynp	111.22	45.43	20
17	Hayden Valley, WY	idas	idas-ynp	110.49	44.68	22
18	Tibb's Butte, WY	idas	n/a	109.45	44.95	20
19	Animas River, CO	idas	n/a	107.57	37.93	13
20	Red Mt. Pass, CO	idas	n/a	107.71	37.90	3
21	Tomboy Road, CO	idas	n/a	107.77	37.94	24
22	Bishop, CA	melissa	melissa-west	118.28	37.17	20
23	Gardnerville, NV	melissa	melissa-west	119.78	38.81	17
24	Red Earth Way, NV	melissa	melissa-west	118.84	38.98	20
25	Silver Lake, NV	melissa	melissa-west	119.93	39.65	18
26	Verdi, NV	melissa	melissa-west	120.00	39.51	20
27	Washoe Lake, NV	melissa	melissa-west	118.82	38.65	8
28	De Beque, CO	melissa	melissa-east	108.21	39.32	20
29	Deeth-Charleston, NV	melissa	melissa-east	115.38	41.30	20
30	Goose Lake, CA	melissa	melissa-east	120.29	41.99	20
31	Lamoille Canyon, NV	melissa	melissa-east	115.47	40.68	19
32	Ophir City, NV	melissa	melissa-east	117.27	38.94	19
33	Star Creek Canyon, NV	melissa	melissa-east	118.12	40.55	16
34	Surprise Valley, CA	melissa	melissa-east	120.10	41.28	20
35	Montague, CA	melissa	melissa-east	122.38	41.77	19
36	Cokeville, WY	melissa	melissa-east	110.94	42.01	10
37	Montrose, CO	melissa	melissa-east	107.82	38.37	20
38	Victor, ID	melissa	melissa-east	111.11	43.66	20
39	Cody, WY	melissa	melissa-rockies	108.98	44.51	23
40	Lander, WY	melissa	melissa-rockies	108.36	42.65	24
41	Yellow Pine, WY	melissa	melissa-rockies	105.40	41.25	19
42	Albion Meadows, UT	melissa	melissa-rockies	111.62	40.59	46
43	Mt. Rose, NV	alpine	mt-rose	119.93	39.32	52
44	Carson Pass, CA	alpine	carson	120.02	38.71	50
45	Corey Peak, NV	alpine	sierra-nevada	118.77	38.45	8
46	Lake Emma, CA	alpine	sierra-nevada	119.48	38.28	32
47	Sonora Pass, CA	alpine	sierra-nevada	119.63	38.33	44
48	Sweetwater Mts., CA	alpine	sierra-nevada	119.33	38.45	23
49	Tioga Crest, CA	alpine	sierra-nevada	119.26	37.97	38
50	South Fork, CA	alpine	white	118.57	37.21	14
51	Reed Flat, CA	alpine	white	118.18	37.38	8
52	County Line Hill, CA	alpine	white	118.19	37.46	40
53	Buck Mt., CA	warner	n/a	120.29	41.69	44
54	Eagle Peak, CA	warner	n/a	120.22	41.26	40

© 2014 John Wiley & Sons Ltd

Table 1 *Continued*

Locality no.	Locality	Taxon	Subgroup	Longitude (°W)	Latitude (°N)	No. of individuals
55	Steens Mt., OR	warner	n/a	118.73	42.66	13
56	Big Ice Cave, WY	jackson	n/a	108.40	45.16	18
57	Hunt Mt., WY	jackson	n/a	107.75	44.68	30
58	Riddle Lake, WY	jackson	n/a	110.55	44.36	28
59	Bull Creek, WY	jackson	n/a	110.55	43.30	44
60	Blacktail Butte, WY	jackson	n/a	110.68	43.64	46
61	Pinnacles, WY	jackson	n/a	109.98	43.74	19
62	Rendezvous Mt., WY	jackson	n/a	110.88	43.60	32
63	Shefield Creek, WY	jackson	n/a	110.66	44.10	26
64	Swift Creek, WY	jackson	n/a	110.91	42.73	4
65	Periodic Spring, WY	jackson	n/a	110.85	42.75	20
66	Dubois, WY	jackson	n/a	109.70	43.56	41

Lycaeides idas occurs in Alaska, throughout Canada, and in the USA Rocky Mountains from Montana and Idaho to Colorado (Nabokov 1943; Scott 1986). Finally, *L. melissa* is found throughout the central and western portions of the USA and southern Canada in montane, steppe and disturbed habitats.

Reproductive isolation between *L. anna*, *L. idas* and *L. melissa* is incomplete (Gompert *et al.* 2010b; Gompert *et al.* 2013b). *Lycaeides anna* and *L. melissa* readily hybridize in the laboratory (M. L. Forister unpublished data), and in previous studies, we documented the existence of several groups of admixed *Lycaeides* populations (Gompert, *et al.*, 2006, 2008; Lucas *et al.* 2008; Gompert *et al.* 2010b; Nice *et al.* 2013). In particular, *Lycaeides* butterflies occur at high elevations in the Sierra Nevada and Sweetwater mountains where they are ecologically distinct from nearby *L. anna* and *L. melissa* populations. Genetic and morphological data suggest that these butterflies are of hybrid origin, and this high-elevation lineage has been recognized as a homoploid hybrid species (Gompert *et al.* 2006; Nice *et al.* 2013). Herein, we refer to this admixed lineage, as well as ecologically similar populations in the White mountains, as 'Alpine *Lycaeides*'. Other similar, but phenotypically more variable, and probably admixed *Lycaeides* butterflies exist in the Warner mountains (hereafter, 'Warner *Lycaeides*'; Gompert *et al.* 2008; Lucas *et al.* 2008; Nice *et al.* 2013). A third series of admixed *Lycaeides* populations occurs in the Rocky Mountains, specifically in the vicinity of Jackson Hole and the Gros Ventre, Snake River, Salt River and Teton mountains (Gompert, *et al.* 2010b, 2012). Butterflies from these populations are ecologically and genetically similar to nearby *L. idas* butterflies, but show evidence of substantial introgression from *L. melissa* (Gompert *et al.* 2010b, 2012 2013a,b). We refer to these butterflies as 'Jackson *Lycaeides*'.

In some analyses, we consider additional subdivisions (subgroups) within nominal *Lycaeides* species that we delineate based on patterns of spatial population structure or variation in genetic ancestry (see 'Spatial analyses of genetic variation and ancestry' in the Results section of this study). In particular, we sometimes distinguished between or among (i) more northern *L. idas* in Montana ('idas-north') and *L. idas* localities on or near the Yellowstone plateau ('idas-ynp') with evidence of more mixed ancestry; (ii) genetically differentiated western, central and eastern *L. melissa* populations; (iii) four spatially disjunct and genetically distinct lineages within Alpine *Lycaeides*, namely the Mt. Rose locality ('mt-rose'), the Carson Pass locality ('carson'), five localities in the central Sierra Nevada or Sweetwater Mts. ('sierra-nevada') and three localities in or near the White Mts. ('white'); and a single Jackson *Lycaeides* locality, Dubois, WY ('dubois'), that included individuals with highly variable admixture proportions perhaps because of very recent or ongoing hybridization (Table 1, Fig. 1).

Data collection

We sampled 1536 adult *Lycaeides* butterflies from 66 localities (Table 1, Fig. 1a). We isolated DNA from each butterfly with Qiagen's DNeasy 96 Blood and Tissue Kit (Cat. No. 69581; Qiagen Inc., Valencia, CA, USA). We then created a reduced complexity, double-digest restriction fragment-based DNA library for each individual using laboratory methods similar to Gompert *et al.* (2012) and Parchman *et al.* (2012). Briefly, we first digested each butterfly's genome with the restriction enzymes *EcoRI* and *MseI*. We then attached adaptor oligonucleotides to the DNA fragments that included the Illumina adaptors and unique 8–10 base pair (bp) identification sequences or barcodes. We PCR-amplified and size-selected the fragment libraries prior to sequencing (see the Supporting information for more details). These DNA libraries were sequenced at the University of Texas Genomic Sequencing and Analysis Facility

© 2014 John Wiley & Sons Ltd

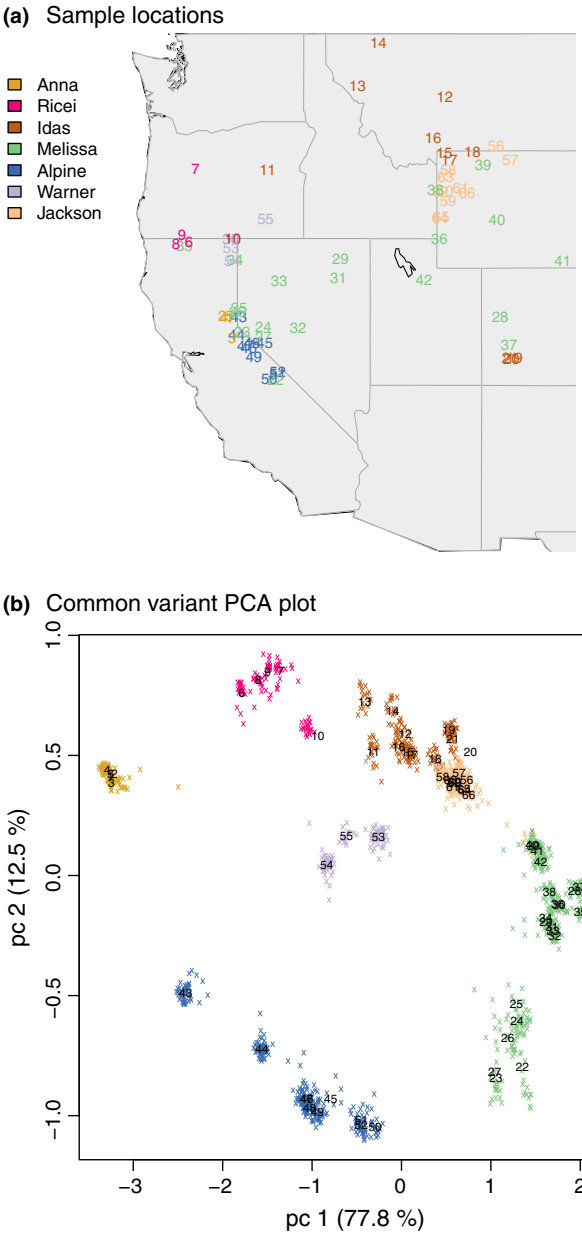

Fig. 1 Sample locations (a) and statistical summary of population-genetic structure based on a principal component analysis of 15 069 common genetic variants (b). The coloured symbols in pane (b) denote individuals' PC scores, and numbers show the mean PC scores for each locality (see Table 1 for locality numbers).

(Austin, TX, USA) on the Illumina HiSeq 2500 platform. We included 384 DNA libraries per lane, and each butterfly's DNA library was sequenced on two lanes. We generated 8 lanes of sequence data or 1.77 billion 100 bp, single-end sequences, with butterflies from multiple taxa and populations included on each lane. We discarded sequences that were similar to the PhiX control or *Wolbachia*, an endoparasitic bacterium that infects many insects, including *Lycaeides* (Gompert *et al.* 2008).

We lack a reference genome for *Lycaeides*. Consequently, we generated a reference sequence set from the Illumina sequence data. We created the reference sequence set by assembling a subset (5 million from each lane; 40 million sequences in all) of the Illumina sequences *de novo* using SEQMAN NGEN ver. 11.0.0.172 (DNASTAR) and extracting the consensus sequences from the high-quality contigs (see the Supporting information for additional information). We then treated each of these 252 000 consensus sequences as an artificial chromosome in a reference-based assembly for each butterfly. Specifically, we mapped each butterfly's sequences onto the reference sequence using BWA ver. 0.7's `aln` and SAMSE algorithms (Li & Durbin 2009). We allowed a 4-bp mismatch between the reference and query sequence, no more than one gap open per query sequence, and only placed query sequences with a unique best match. We used SAMTOOLS ver. 0.1.18 to index, sort and merge the individual alignments.

Estimation of genetic variation

We estimated the folded-site allele frequency spectrum, nucleotide diversity (π) and θ to describe genetic variation within each butterfly population. We obtained maximum-likelihood estimates of these parameters using the expectation-maximization (EM) algorithm described by Li (2011). This method does not require genotype or variant calling, but instead maximizes the model likelihood with respect to the genotype likelihood. We used SAMTOOLS ver. 0.1.18 for this analysis and iterated the EM algorithm 20 times for each population to ensure numerical convergence.

Next, we identified single nucleotide variants (SNVs) from the sequence data by calculating the Bayesian posterior probability that each nucleotide was variable with SAMTOOLS and BCFTOOLS ver. 0.1.18. When identifying variants, we used the full prior with θ set to 0.001, and we ignored aligned sequences with phred-scaled mapping quality <20 and bases with phred-scaled base quality <13. We only designated genetic variants at nucleotide sites where we had sequence data for at least 80% of the sampled butterflies and where the posterior probability of the sequence data under a null model that the nucleotide was invariant was <0.01. We estimated global variant MAFs directly from genotype likelihoods using the maximum-likelihood approach described by Li (2011) and used these estimates to assign loci to MAF classes. We identified 801 218 SNVs based on these criteria, which included 28 701 common variants (MAF ≥ 5%), 180 043 low-frequency variants

© 2014 John Wiley & Sons Ltd

(0.5% ≤ MAF < 5%) and 349 497 rare variants (0.1% ≤ MAF < 0.5%). We ignored very rare variants (i.e. MAF < 0.1%) as these occurred in only one or two individuals. The median sequencing depth per individual per variable site was 7.6 reads (mean = 8.5 reads, standard deviation = 4.5 reads).

Spatial analyses of genetic variation and ancestry

As a first step to quantify the spatial distribution of genomic variation, we developed and implemented a statistical model to jointly estimate genotypes and admixture proportions from common, low-frequency and rare SNVs. Our admixture proportion model is similar to the correlated allele frequencies admixture model in STRUCTURE (Pritchard *et al.* 2000; Falush *et al.* 2003). The most important difference is that we incorporate sequence coverage, sequence error and alignment errors in our model. A similar approach was proposed independently by Skotte *et al.* (2013) and shown to decrease bias relative to analysing called genotypes. As in the STRUCTURE admixture model, we assume each individual's genome is potentially a mosaic or mixture of genetic loci inherited from each of K source populations. A set of admixture proportions, denoted $q_1, q_2, \ldots, q_K$, indicate the proportion of an individual's genome that was inherited from each source population (these parameters quantify global or genome-average genetic ancestry; Gompert & Buerkle 2013). Thus, the key parameters in this admixture model are the unknown genotypes, source population allele frequencies and admixture proportions. We describe this model in more detail and provide an evaluation of this method with simulated data in the Supporting information.

We estimated these model parameters in a Bayesian framework using Markov chain Monte Carlo (MCMC) and used deviance information criterion (DIC) to compare models with different values of K. We analysed common, low-frequency and rare variants separately. We considered only one SNV per reference consensus sequence to reduce the effect of nonindependence among physically linked loci (this does not preclude possible linkage disequilibrium between SNVs from different reference sequences). 15 069 common variants met this criterion, and we randomly selected 15 069 each of the low-frequency and rare variants to facilitate comparisons among data sets. Thus, the results are based on 15 069 common, low-frequency and rare SNVs (45 207 SNVs total). We generated samples from the posterior probability distribution of the model parameters with one to eight assumed source populations. We ran two independent MCMC analyses with 15 000 iteration chains, sampled every fifth iteration and preceded by 5000 iteration burn-ins, for each model. All parameter estimates were based on samples from both chains.

We used principal component analysis (PCA) to statistically summarize the distribution of genotypic variation across the sampled *Lycaeides* butterflies. We first calculated the expected genotypic value for each individual and locus from the results of the admixture model as the average of expectations from models with 2–8 source populations (we did not use the result with one source population, as this model did not fit the data well). With that said, genotype posterior probabilities from models with different numbers of source populations were highly correlated ($r \geq 0.98$ in all cases and $r \geq 0.99$ when excluding $k = 1$), and thus, averaging over models had a negligible effect on genotype estimates. Our average genotype estimates were also highly correlated with estimates obtained with a uniform prior on genotype ($r = 0.93$). The genotype estimates ranged from 0 to 2 and are not constrained to integer values, but rather denote the best estimate of the number of nonreference allele copies for an individual and locus (see the Supporting information for additional details; Gompert *et al.* 2013a). We used these genotype estimates to calculate a separate genetic covariance matrix based on common, low-frequency and rare variants, and we summarized these genetic variance matrices with PCA (the genetic covariance matrix is a N by N symmetric matrix with the genetic covariance between individuals in off-diagonal elements and the genetic variance within individuals in the diagonal elements; Price *et al.* 2006). We computed genetic covariance matrices and performed PCA in R using the PRCOMP function using centred, but not standardized, genotypes (R Core Team 2013).

We quantified spatial genetic structure for common, low-frequency and rare variants based on the spatial autocorrelation in sample allele frequencies between localities to quantify isolation by distance. We calculated the sample allele frequency for each population and locus by summing the expected genotypic values and dividing by twice the number of sampled individuals. We then estimated Moran's I to summarize the allele frequency correlation between pairs of populations separated by different geographic distances (we used 25 and 75 km distance bins; Epperson 2003). We averaged Moran's I across loci following Eckert *et al.* (2010) and tested for significant deviations from random spatial patterns in each distance class using a randomization test with 1000 permutations of population labels. We wrote the functions for these analyses in R.

We then used Bayesian regression analysis to answer our first focal question and determine whether genetic differences between butterfly populations were best explained by geographic distances, nominal taxon

© 2014 John Wiley & Sons Ltd

boundaries or both. The regression analysis was based on the mixed model framework proposed by Clarke *et al.* (2002) to account for the correlated error structure inherent in pairwise observations such as genetic distances. We assumed that $\text{logit}(y_{ij}) = \beta_0 + \beta_{geo}X_{ij}^{geo} + \beta_{taxon}X_{ij}^{taxon} + \lambda_i + \lambda_j + \epsilon_{ij}$, where y_{ij}, X_{ij}^{geo} and X_{ij}^{taxon} denote the genetic (Nei's D_A; Nei *et al.* 1983; Takezaki & Nei 1996), geographic distance (great circle) and taxonomic identity (1 for different taxa vs. 0 for the same taxon) between populations i and j, the β's are fixed-effect regression coefficients, the λ's are random effects representing the average deviation of genetic distances involving localities i or j from the expectation based on their geographic and taxonomic distances, and ε_{ij} is the residual error. We centred (both) and standardized (geographic distance only) covariates prior to analysis. We specified uninformative Gaussian priors for the regression coefficients ($\mu = 0$, $\sigma^2 = 1000$), a hierarchical Gaussian prior for the population random effects ($\mu = 0$, $\sigma^2 = \sigma^2_{pop}$) and uninformative gamma priors for the reciprocals of the random effect and residual variances ($\alpha = 1$, $\beta = 0.01$). We fit the full model and reduced models for common, low-frequency and rare variants in R via the RJAGS interface with MCMC models in JAGS (Plummer 2003). We compared models that included geographic distances, taxonomic distances or both using DIC. In each case, we ran three independent MCMC chains each with 10 000 iterations, a 2000 iteration burnin and a thinning interval of five. We calculated Gelman and Rubin's scale reduction factor to verify adequate chain mixing and probably convergence to the posterior distribution (Gelman & Rubin 1992; Plummer *et al.* 2006).

Tests of the bifurcating tree model

We conducted two sets of analyses to answer our second focal question and formally test whether the evolutionary history of *Lycaeides* populations is consistent with a bifurcating tree. We first used the TREEMIX method proposed by Pickrell & Pritchard (2012) to infer a population graph representing the hypothesized evolutionary history of the sampled *Lycaeides* populations. Population graphs allow both population splits and admixture or migration events. We constructed *Lycaeides* population graphs based on the matrix of allele frequency covariance between pairs of populations. We only considered common variants, as we found little evidence of mixed ancestry in admixture proportion estimates from low-frequency or rare variants (see Results). We rooted the population graph with *L. anna* and constructed graphs allowing 0–10 admixture events. We calculated the proportion of the variance in population covariances explained by the population graph with different numbers of admixture events to quantify model fit, and we used a block-jackknife procedure with blocks of 10 SNVs to determine whether individual admixture events significantly improved model fit (we considered blocks of 10 SNVs for computational convenience; Pickrell & Pritchard 2012). A complementary analysis of the null-bifurcating tree hypothesis based on 3-population or f_3-statistic test (Reich *et al.* 2009; Patterson *et al.* 2012) is described in the Supporting information.

Distinguishing between contemporary and historical admixture

We developed and implemented an admixture class model to quantify the genome composition of butterflies with mixed ancestry and thereby answer our third focal question. Although estimates of genome-average ancestry, **q**, provide evidence of admixture, they do not provide information about how patterns of ancestry are organized across the genome. However, estimates of admixture class frequencies provide estimates of how much of the genome is heterozygous for ancestry (inter-source ancestry) versus homozygous for ancestry (intra-source ancestry). This distinction is important for discriminating between hybrids with recent nonadmixed parents (which should have high inter-source ancestry) and late generation hybrids (with low inter-source ancestry) that would be expected in a stable hybrid lineage (e.g. Buerkle & Rieseberg 2008; Gravel 2012), and provides additional information about demographic and evolutionary dynamics in admixed populations. For example, individuals with one or two nonadmixed parents (back-cross or F1 individuals) are expected to have higher inter-source population ancestry than individuals with similar admixture proportions with only admixed parents (similar logic is used to assign individuals to hybrid classes in the admixture model in NEWHYBRIDS; Anderson & Thompson 2002). To construct this model, we replaced the admixture proportion vector **q** with an admixture class matrix **Q**, where, for example, Q_{11} denotes the proportion of an individual's genome where both gene copies are descended from source population 1, Q_{22} denotes the proportion of an individual's genome where both gene copies are descended from source population 2 and $Q_{12} = Q_{21}$ denotes the proportion of an individual's genome where one gene copy is descended from source population 1 and the other from source population 2. That is, Q_{12} quantifies inter-source population ancestry. We only consider the case of $k = 2$. Aside from substituting **q** with **Q** to describe genome-average ancestry and serve as the prior for locus-specific ancestry, this

© 2014 John Wiley & Sons Ltd

model is identical to the admixture proportion model we described above (see 'Spatial analyses of genetic variation and ancestry' in the Methods section) and very similar to the correlated allele frequency admixture model in STRUCTURE (Falush *et al.* 2003). We demonstrate the efficacy of this model by applying it to simulated data (see the Supporting information for a description of the simulations and the results).

We used the admixture class model to estimate admixture classes (**Q**) based on common genetic variants. We separately analysed four sets of *Lycaeides* entities: (i) *L. anna*, 'melissa-west' and Alpine *Lycaeides*, (ii) *L. anna*, 'melissa-east' and Warner *Lycaeides*, (iii) *L. anna*, 'melissa-east' and *L. idas* and (iv) 'melissa-east' and 'melissa-rockies', *L. idas* and Jackson *Lycaeides* (Table 1). Each set includes a putatively admixed lineage (listed last in each of these sets) and possible source populations based on geography and previous investigations (e.g. PCA and admixture proportion estimates; see Results). These analyses assumed two populations ($k = 2$). We ran two independent MCMC analyses with 15 000 iteration chains, sampled every fifth iteration and preceded by 5000 iteration burn-ins, for each set of individuals. We estimated admixture classes using samples from both chains. We then fit linear random-effect models of admixture class as a function of locality to quantitatively summarize variation in genome composition within and among localities. We used restricted maximum likelihood to estimate variance components with the LMER function in the R package LME4 (Bates *et al.* 2013; R Core Team 2013).

Results

Estimation of genetic variation

Genome-wide measures of genetic diversity for all nucleotides were similar among populations (mean $\pi = 0.0051$, SD = 0.0008; mean $\theta_w = 0.0046$, SD = 0.0002; Fig. 2). Maximum-likelihood estimates of the folded allele frequency spectrum indicated that most nucleotides were invariant or had a rare segregating minor allele. Consistent with these results, we confidently identified many more rare (349 497) and low-frequency (180 043) variants than common variants (28 701) segregating in the sampled *Lycaeides* butterflies.

Spatial analyses of genetic variation and ancestry

Genetic variation at common variants was best explained by an admixture model with four source populations, whereas models with larger numbers of source populations better explained variation at rare and low-frequency variants (Fig. S1a, Supporting information). Similarly, most (90.3%) of the variation in common variant genetic covariance among pairs of individuals was accounted for by the first two principal components, but the first ten principal components only explained 17.3% of the rare variant genetic covariance (Fig. S1b, Supporting information). The first two principal components for common variants separated named entities and individuals from different localities (Fig. 1b). *L. anna* and *L. melissa* individuals were most differentiated based on principal component (PC) 1, whereas putatively admixed lineages had intermediate PC 1 scores and intermediate or extreme PC 2 scores. The nominal taxa *L. anna ricei* and *L. idas* also had intermediate PC 1 scores and had similar PC 2 scores. Principal component analysis based on common genetic variants also identified geographic population structure within several entities, namely *L. melissa* and Alpine *Lycaeides*. Principal components of low-frequency variants did not separate entities or populations as well, but showed patterns of population structure that were qualitatively consistent with those based on common variants (Fig. S2, Supporting information). In contrast, one or several localities were distinguished from most other entities and localities on each principal component for rare variants (Fig. S3 vs. Fig. S4, Supporting information).

We found evidence of isolation by distance in *Lycaeides* based on estimates of Moran's I for common, low-frequency and rare variants, with positive allele frequency correlations between localities separated by shorter geographic distances (i.e. <400 km) and negative allele frequency correlations at greater distances (i.e. beyond 800 km). But, we detected more substantial positive and negative spatial auto-correlation for common variants than rare variants (Fig. 3a). For example, allele frequency correlations between pairs of populations within 25 km of each other were 0.64, 0.37 and 0.15 for common, low-frequency and rare variants, respectively. Consistent with these results, Bayesian regression analysis indicated that more distant populations or those classified as different taxa were genetically more distinct (Fig. 3b; posterior means and 95% equal-tail probability intervals: common variants, $\beta^{geo} = 1.09$ (1.05−1.14), $\beta^{taxon} = 0.256$ (0.238−0.273); low-frequency variants, $\beta^{geo} = 0.509$ (0.485−0.534), $\beta^{taxon} = 0.166$ (0.157−0.176); and rare variants, $\beta^{geo} = 0.244$ (0.232−0.255), $\beta^{taxon} = 0.0765$ (0.0718−0.0811). The full model with geographic and taxonomic distance received the most support for all variant classes based on DIC, but in each case, the reduced model with only taxonomic distances was preferred over the model with only geographic distances (Table S1, Supporting information).

Admixture proportions inferred from common variants indicated that many individuals had mixed

© 2014 John Wiley & Sons Ltd

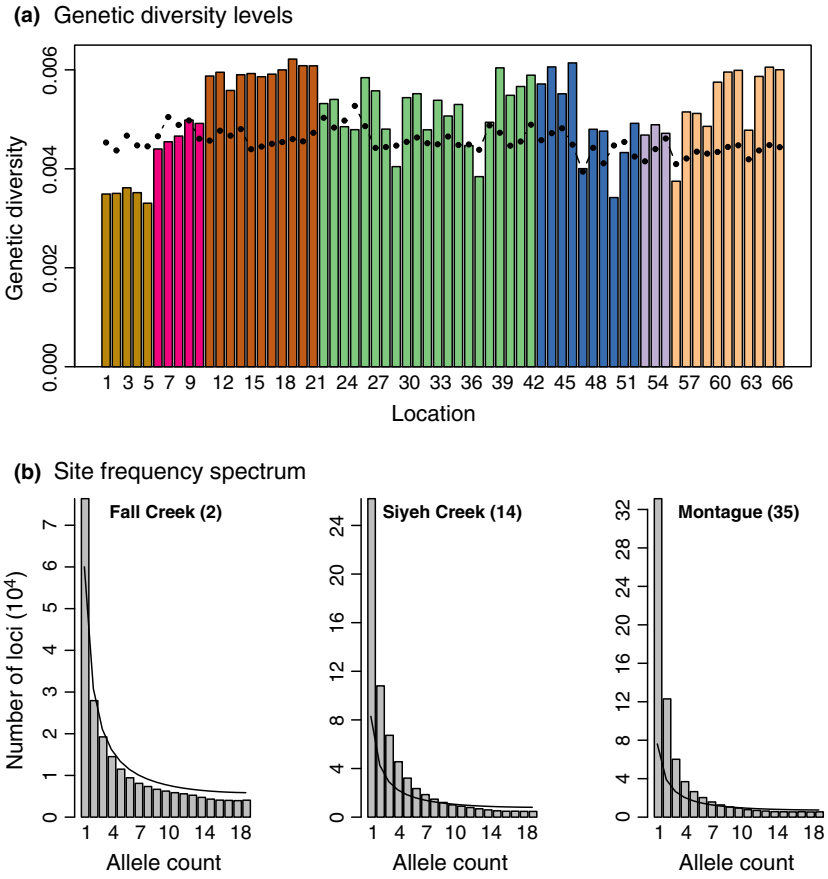

Fig. 2 Genetic diversity within sample locations. Pane (a) depicts genome-wide estimates of expected heterozygosity (π; bars) and Watterson's θ (points) for each location (locality numbers are defined in Table 1). Bars are coloured by taxa as defined in Fig. 1. The genome-wide folded site allele frequency spectrum in three of these localities is shown in pane (b) (the plots exclude invariant loci). Null expectations from the standard neutral model are indicated by a solid line. Genetic diversity estimates were based on an average of 16.2 million nucleotides per population (SD = 1.9 million).

ancestry regardless of the number of source populations assumed (Fig. 4). For example, when we assumed 2–5 source populations, Warner *Lycaeides* admixture proportions (**q**) showed that these individuals had genetic ancestry from two or three hypothetical source populations that were genetically similar to extant *L. anna*, *L. anna ricei*, *L. melissa*, Alpine *Lycaeides* and perhaps *L. idas*. Although all Jackson *Lycaeides* individuals had mixed ancestry when 2–4 source populations were assumed, multiple individuals from a single locality (locality number 66, Dubois, WY, USA) had mixed and variable ancestry even when a greater number of source populations were assumed. We also documented mixed ancestry in *L. idas* individuals, which were not previously hypothesized to be admixed. Finally, consistent with the PCA results, we identified spatial variation in genetic ancestry among localities for some species or entities, including *L. melissa* and Alpine *Lycaeides*. In contrast, we found little evidence of mixed ancestry based on rare or low-frequency variants (Figs 5 and S5, Supporting information). Rather, based on rare or low-frequency variants, most individuals and localities had genetic ancestry principally from one source population.

Tests of the bifurcating tree model

The rooted population tree without admixture explained 87% of the variance in population covariances, but this increased as admixture events were added to the graph (Fig. S6, Supporting information). Here, we focus on the graph with seven admixture events (Fig. 6), which explained 96% of the variance in population covariances. All seven migration events significantly improved the fit of the model to the data ($P < 10^{-30}$). This population graph indicated that admixture occurred between *L. anna* and *L. melissa* around the origin of Alpine *Lycaeides* with subsequent admixture events in specific Alpine populations. We also found evidence of separate admixture events in the

© 2014 John Wiley & Sons Ltd

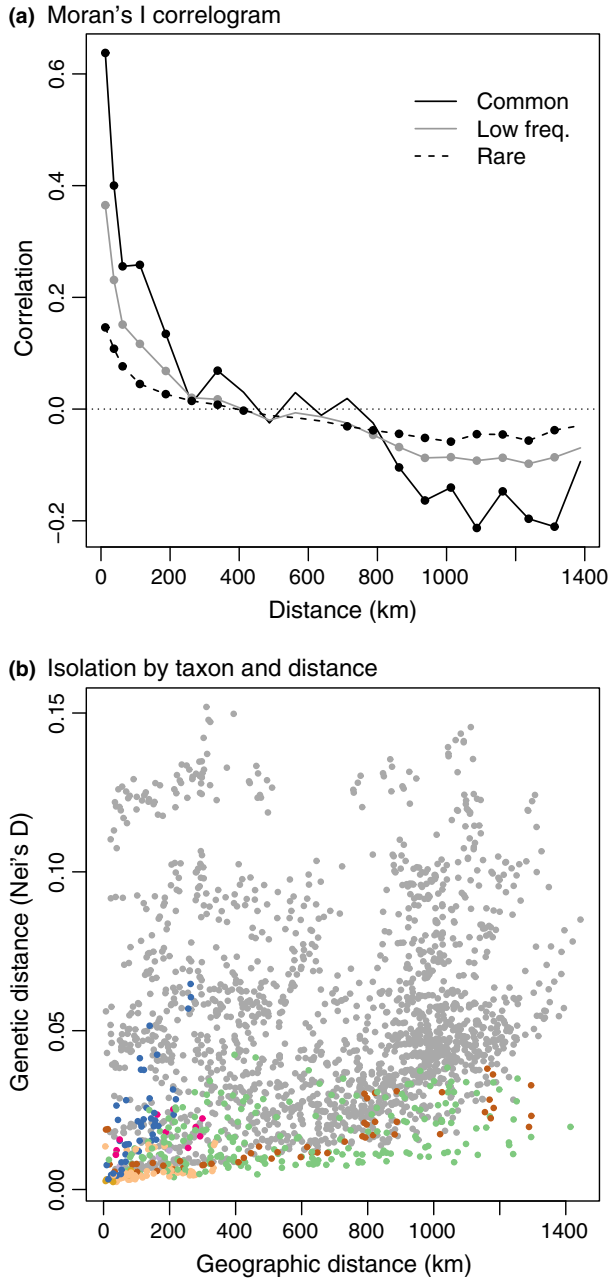

Fig. 3 Plots of isolation by distance. Moran's I correlogram depicts the correlation of allele frequencies between locations as a function of distance (a). Solid circles indicate correlations that are significantly different (permutation $P<0.05$) than expected under the null hypothesis of random spatial distribution of the sample allele frequencies. The scatterplot in pane (b) shows the relationship between geographic and genetic distance (Nei's D for common variants) for pairs of populations with the same (coloured circles) or different (grey circles) taxonomic designations (different colours denote different taxa as defined in Fig. 1).

Mt. Rose Alpine *Lycaeides* population (locality 43) and one of the Warner *Lycaeides* populations (Steens Mt., locality 55), as well as gene flow from Alpine *Lycaeides* back into *L. melissa*. We obtained similar results using alternative taxa to root the population graph (Fig. S7, Supporting information).

Distinguishing between contemporary and historical admixture

Admixture class (**Q**) estimates indicated that Alpine *Lycaeides* (including all four subgroups), Warner *Lycaeides*, Jackson *Lycaeides* and *L. idas* individuals' genomes were a mosaic of ancestry from different source populations. Specifically, these individuals' genomes were composed of genetic regions where (i) both gene copies were inherited from source population one (nonzero Q_{11}), (ii) both gene copies were inherited from an alternative source population (nonzero Q_{22}) and (iii) each gene copy was inherited from a different source population (nonzero Q_{12}; Fig. S8, Supporting information). We found less evidence of inter-source population ancestry (i.e. nonzero Q_{12}) in *L. idas* than in the other admixed lineages. And we found very few nonadmixed individuals in the 'melissa-east' and 'melissa-rockies', *L. idas*, and Jackson *Lycaeides* population set (Fig. S8, Supporting information).

Whereas many admixed individuals had inter-source population ancestry, levels of inter-source population ancestry were less than predicted if source populations were mating randomly in a hybrid swarm (here $E[Q_{12}] = 2q_1[1-q_1]$; Fig. 7). We also identified a small number (1–3) of admixed individuals in populations of putatively 'pure' *L. anna*, *L. idas* and *L. melissa*. These individuals showed evidence of mixed ancestry and estimates of inter-source population ancestry that were approximately as high as possible given their global genetic ancestry (q_1; Figs 7 and S8, Supporting information). Thus, these individuals were likely the offspring of back-crosses (one parent with ancestry in one of the source populations and one hybrid parent), and thereby suggest contemporary hybridization in the parents of the sampled individuals.

We detected more variation in genetic ancestry and admixture classes among individuals in some entities than others. In particular, Alpine *Lycaeides* butterflies varied substantially in their genome composition (Figs 7 and S8, Supporting information). Most of the variation in global genetic ancestry (q_1) occurred among localities in Alpine *Lycaeides* (90.0%), Warner *Lycaeides* (82.1%) and *L. idas* (75.5%), but not Jackson *Lycaeides* (43.1%; Fig. S9, Supporting information). Conversely, inter-source population ancestry (Q_{12}) varied more within localities than among localities in most entities (percentage of variance among localities: Warner = 37.1%, Jackson = 20.1%, *L. idas* = 25.2%), but not in Alpine *Lycaeides* (59.2%; Fig. S10, Supporting information).

© 2014 John Wiley & Sons Ltd

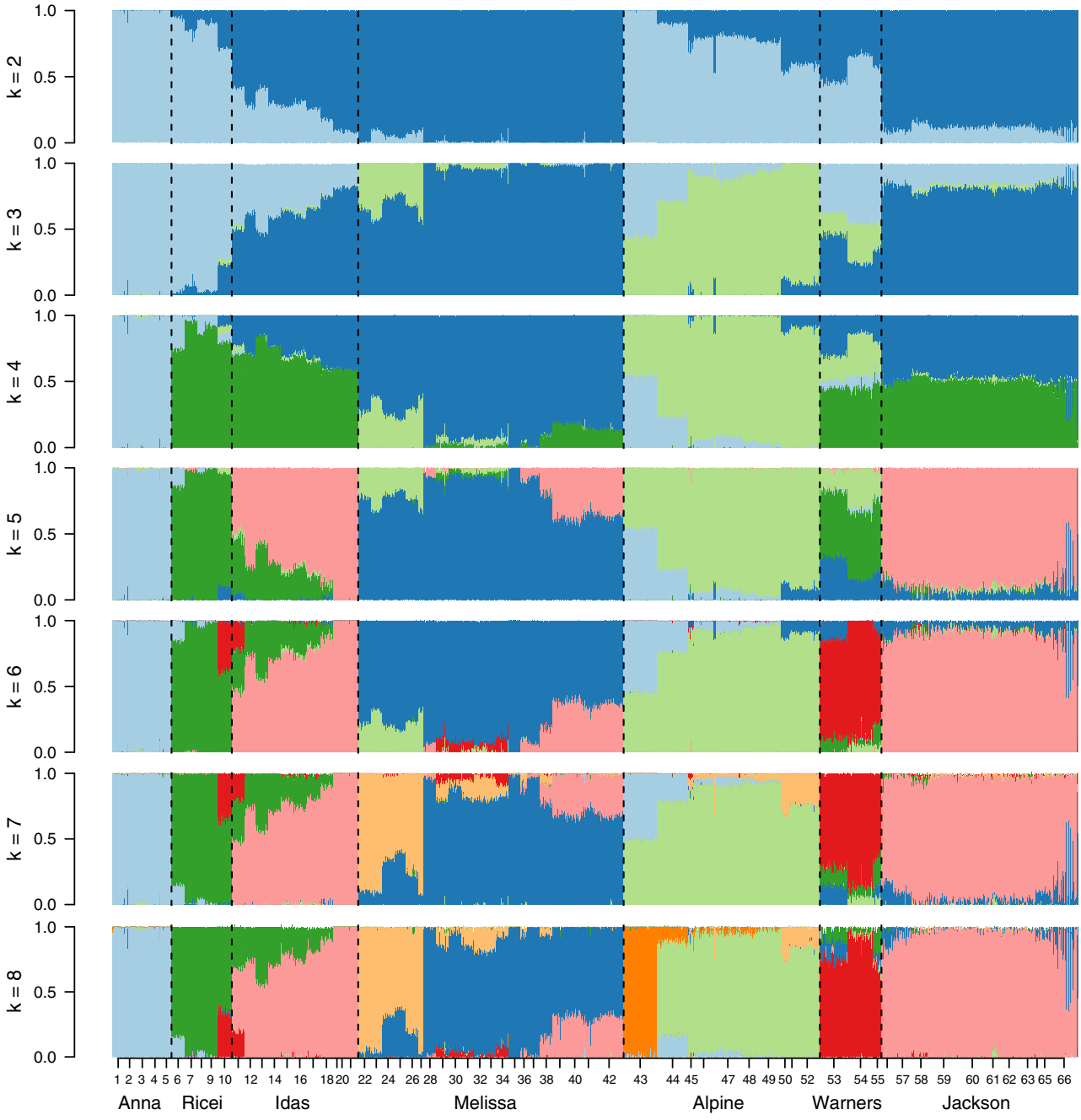

Fig. 4 Admixture proportions based on common genetic variants. Each bar corresponds to an individual, and each bar is coloured to depict the Bayesian point estimates of the individual's admixture proportions. That is to say, each coloured segment depicts the proportion of an individual's genome inherited from one of k inferred source populations. Results with 2–8 putative source populations are shown (k = 4 was the best model based on DIC, Fig. S1, Supporting information). Tick marks and numbers below the plots identify localities based on the locality numbers in Table 1.

Discussion

We found that the evolutionary history of *Lycaeides* butterflies is not well-described by a strictly bifurcating tree; instead, genomic variation in this group has been structured by spatial isolation within nominal species and admixture and introgression between them (tests of admixture are summarized in Table 2). In particular, tree- and genetic ancestry-based analyses revealed patterns of genomic variation consistent with geographically widespread historical admixture and introgression in cases where this was previously known or suspected

© 2014 John Wiley & Sons Ltd

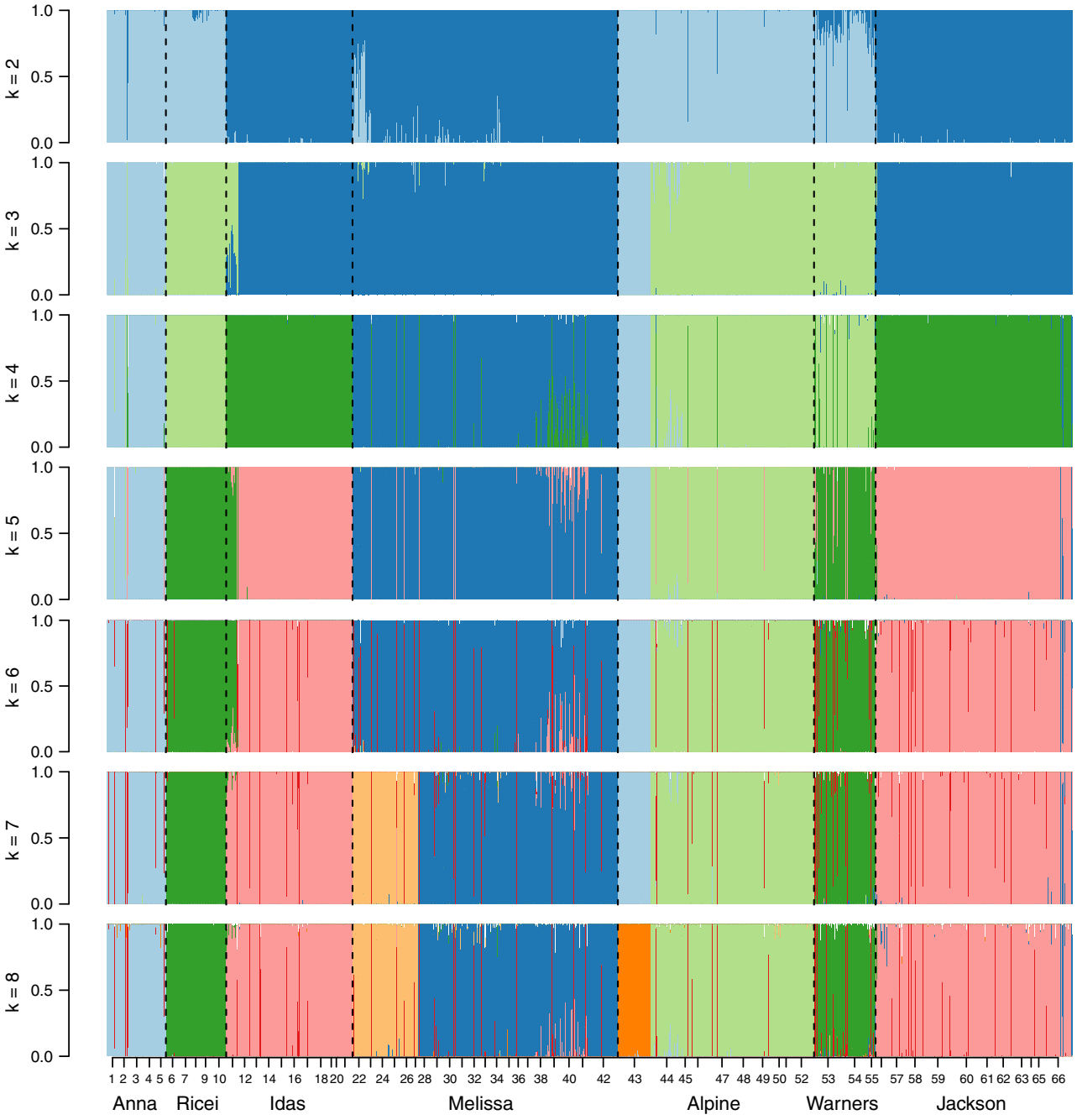

Fig. 5 Admixture proportions based on rare genetic variants. Each bar corresponds to an individual, and each bar is coloured to depict the Bayesian point estimates of the individual's admixture proportions. That is to say, each coloured segment depicts the proportion of an individual's genome inherited from one of k inferred source populations. Results with 2–8 putative source populations are shown (k = 8 was the best model based on DIC, Fig. S1, Supporting information). Tick marks and numbers below the plots identify localities based on the locality numbers in Table 1.

from more limited data (e.g. Alpine and Jackson *Lycaeides*; Gompert, *et al*., 2006, 2012; Nice *et al*. 2013) and in cases where admixture was not thought to have occurred (e.g. in *L. idas*). Analyses of genetic ancestry indicated that the genome composition of individuals with mixed ancestry varied little within populations, consistent with the hypothesis that these admixed populations are independent evolutionary lineages rather than active hybrid zones. In contrast, genetic ancestry differed among conspecific populations, particularly in

© 2014 John Wiley & Sons Ltd

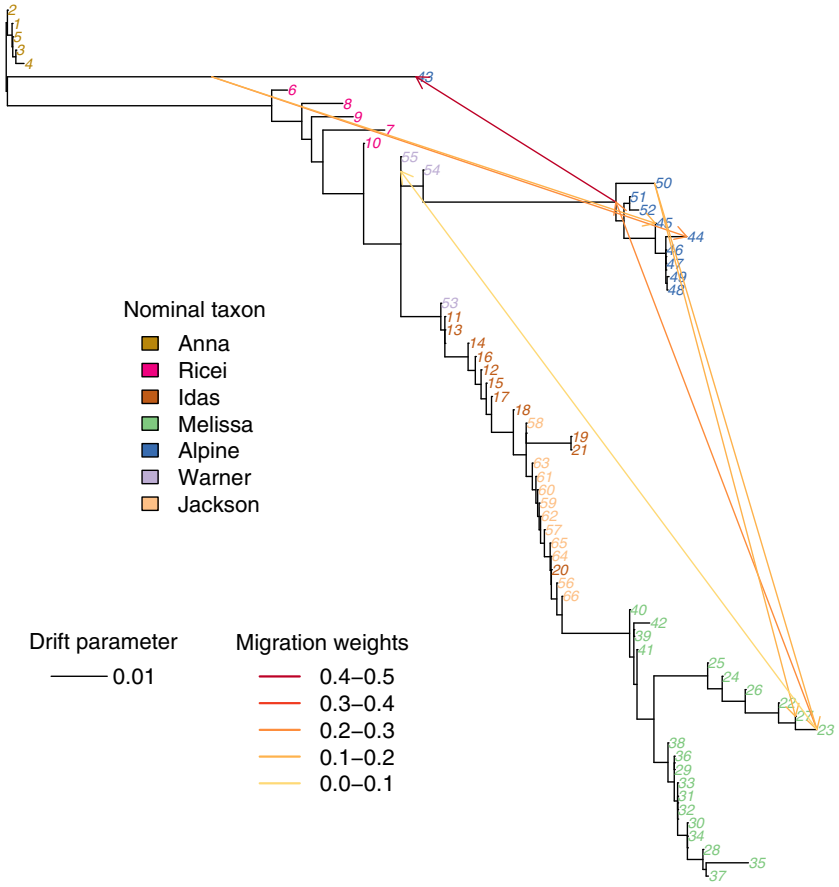

Fig. 6 Population graphs inferred by TREEMIX for *Lycaeides* butterflies from 66 localities, allowing seven migration or admixture events. Terminal nodes are labelled by locality number (see Table 1) and coloured according to nominal lineage or taxon. Branch lengths are proportional to the evolutionary change (the drift parameter, which is closely related to F_{ST}) along each branch, and migration arrows are coloured according to the migration weight (approximately, the proportion of genetic ancestry from the immigrant population). All seven migration events significantly improve the model ($P < 10^{-30}$).

Alpine *Lycaeides*. These differences in genetic ancestry exist despite similar ecologies, as these butterflies occur in similar high-elevation habitat and use the same alpine-endemic host plant (Gompert *et al.* 2006; Nice *et al.* 2013). We also identified recent hybrids, but only in a few localities. Finally, we found similar patterns of genetic isolation by taxon and isolation by distance with common, low-frequency and rare genetic variants, but evidence of admixture came mostly from common variants. We discuss and interpret these findings in more detail below.

Hybridization and the structure of biological diversity in Lycaeides butterflies

We documented mixed ancestry in Alpine *Lycaeides*, Warner *Lycaeides*, Jackson *Lycaeides* and *L. idas* based on sequence data from 15 069 common genetic variants. Consistent with previous work (Gompert *et al.* 2006; Nice *et al.* 2013), the results indicate that Alpine and Warner *Lycaeides* are admixed and genetically distinct from other *Lycaeides* (Figs 1, 6 and 7). Coupled with previously published morphological and ecological data (Fordyce *et al.* 2002; Nice *et al.* 2002; Fordyce & Nice 2003; Lucas *et al.* 2008; Nice *et al.* 2013), we interpret this as evidence that these entities are isolated lineages of hybrid origin. We have reached this conclusion before, but with fewer genetic markers and localities and without individual or locality-level data (Gompert *et al.* 2006; Nice *et al.* 2013). Results consistent with geographically widespread admixture in *L. idas* were unexpected and likely would have been missed with more limited taxonomic sampling. With that said, negative f_3 for this species could be caused by isolation by distance at the species complex level. Moreover, we do not know how the inclusion of *L. idas* from further north in Canada or Alaska might have affected this result. We also documented possible admixture in the 'melissa-rockies'

© 2014 John Wiley & Sons Ltd

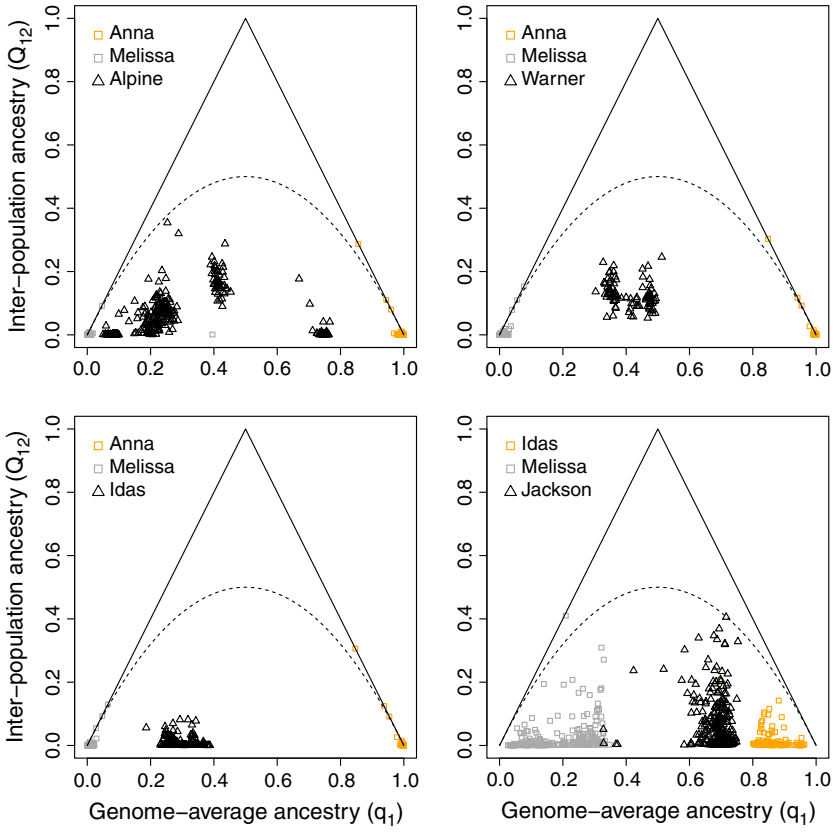

Fig. 7 Scatterplots show the relationship between global genetic ancestry (q_1) and inter-source population ancestry (Q_{12}). Symbols correspond to individuals. Lines indicate the maximum possible inter-source population ancestry given global genetic ancestry (solid lines; individuals on these line have at least one nonadmixed parent) or the expected inter-source population if all individuals mated randomly (dashed line; this corresponds to $2q_1[1-q_1]$).

subgroup, which consists of localities in close spatial proximity to *L. idas* and Jackson *Lycaeides* localities. Finally, we documented a few recent hybrids in *L. anna* and *L. melissa* populations (putative offspring from back-crosses; Figs 7 and S8, Supporting information) and a mixture of individuals with very different genetic ancestry in one Jackson *Lycaeides* population (locality 66 or Dubois; Figs 4 and S8, Supporting information). Thus, opportunities for interspecific gene flow persist in this species complex.

The results show that multiple differentiated lineages exist within Alpine *Lycaeides* (Figs 1 and S8, Supporting information). These lineages could be the product of multiple, independent admixture events, a single admixture event followed by spatial isolation and population differentiation prior to the fixation of different ancestry segments in different populations or some combination of these (nonzero estimates of inter-source population ancestry indicate that this is an ongoing process; Figs 6, 7, and S8, S10, Supporting information). In contrast, the genome composition of Jackson *Lycaeides* approaches that of nearby *L. idas* individuals on the Yellowstone plateau (Figs 1 and 7). Combined with clear evidence of current hybridization at the Dubois locality (locality 66; Fig. S8, Supporting information) and limited ecological differentiation from *L. idas* (Gompert, *et al.* 2010b, 2013b), we interpret this limited genetic discontinuity as evidence of more recent historical admixture or introgression in Jackson *Lycaeides*. In general, the prevalence of hybrid individuals and admixed populations suggests that hybridization in *Lycaeides* could fuel adaptation and diversification from standing genetic variation (e.g. Seehausen 2004; Barrett & Schluter 2008), as has been proposed for other organisms, including European aspen (de Carvalho *et al.* 2010), *Cottus* sculpins (Nolte *et al.* 2005), *Catostomus* suckers (McDonald *et al.* 2008) and *Heliconius* butterflies (Dasmahapatra *et al.* 2012).

Although we generally obtained similar results with different analyses, there were a few exceptions (Table 2). In particular, the population graph and f_3 statistics support the hypothesis that Alpine *Lycaeides* were admixed, but f_3 statistics only indicate admixture in one of the four genetically distinct lineages (subgroups) within this entity ('carson'; (Tables S2 and S3,

© 2014 John Wiley & Sons Ltd

Table 2 Summary of evidence for and against ancient admixture, recent admixture (i.e. early-generation hybrids), repeated admixture (i.e. multiple admixture events), substantial population structure with an entity and the evolutionary independence of each putative admixed lineage. Analyses that support each inference are in bold, whereas analyses that were performed but did not support an inference are shown in plain text. This summary includes results from this study and from previous manuscripts (the latter are indicated with an asterisk): amx = admixture proportion and admixture class estimates; tmx = TREEMIX; f_3 = 3-population tests; abc = approximate Bayesian computation model selection (Nice *et al.* 2013); mcl = analyses of morphological clines (Gompert *et al.* 2010b); gen = PCA of genetic covariance matrix; hpp = unique host plant preference (Fordyce *et al.* 2011; Gompert *et al.* 2013b); eco = presence of novel ecologically important traits (Fordyce & Nice 2003; Gompert *et al.* 2006).

Inference	Alpine	Warner	Jackson
Ancient admixture	**amx**, **tmx**, f_3, **abc***	**amx**, **tmx**, **f_3**, **abc***	**amx**, tmx, **f_3**, **mcl***
Recent admixture	amx	amx	**amx**
Repeated admixture	**tmx**	tmx	tmx
Structured	**amx**	amx	amx
Evo. independence	**gen**, **hpp***, **eco***	**gen**, **hpp***	gen, hpp*

Supporting information). In contrast, TREEMIX and admixture class estimates indicate that all Alpine subgroups were admixed (Figs 6 and S8, Supporting information). Substantial evolutionary change within an admixed population can result in positive (i.e. nonsignificant) f_3 statistics despite admixture, and we hypothesize that this is the cause of this discrepancy. Also, unlike the other methods we used, TREEMIX did not indicate admixture in Jackson *Lycaeides*. We suspect this is because of the genetic similarity between *L. idas* and Jackson *Lycaeides*. Finally, the validity of inferences from each analysis depends on the appropriateness of the model, which means we could be misled by an analysis if our assumptions are incorrect. Thus, whereas we are more confident about conclusions supported by multiple analyses (Table 2), healthy scepticism is warranted even in these cases.

Differences between common and rare variants

Rare and low-frequency variants were much more prevalent than common variants in *Lycaeides* butterflies and were shared among individuals within a population but had limited spatial distributions (Figs 3 and S3, Supporting information). Under the standard neutral coalescent, we expect most variants to be rare, and other factors including population growth and purifying selection can further increase the proportion of variants that are rare (e.g. Williamson *et al.* 2005; Nelson *et al.* 2012). Rare variants could be recent mutations that have not yet spread among populations or older variants that have been present long enough to reach a 'quasi-equilibrium' under migration and genetic drift (Slatkin 1985; Slatkin & Takahata 1985; Barton & Slatkin 1986). Thus, regardless of whether rare variants are young or old, one would expect them to be spatially restricted if dispersal and gene flow are limited (Barton & Slatkin 1986). This means that rare variants should reveal fine-scale spatial genetic structure distinguishing individuals from one or a few localities, as we observed in *Lycaeides* butterflies (Figs 3 and S3, Supporting information). Measures of recent coancestry based on haplotype blocks have similarly been shown to detect fine-scale spatial genetic structure (Lawson *et al.* 2012), and the same is true, albeit to a lesser extent, for rapidly evolving microsatellites (Balloux & Lugon-Moulin 2002; Girod *et al.* 2011; Molfetti *et al.* 2013). The average age of rare variants could have other effects on their geographic distribution. In particular, we would not expect rare variants to provide evidence of historical events that occurred before the mutations at these loci arose, which could explain the lack of signal for admixture from rare variants in *Lycaeides*.

We think that these notable differences in the spatial distribution and history of common and rare variants have possible general implications for studies of historical demography, trait genetics and the genetics of adaptation and speciation. First, as we discussed in the previous paragraph, historical inferences from rare variants might be limited to recent events and population dynamics. This is important as rare variants should constitute the majority of data in modern sequencing studies (e.g. Gravel *et al.* 2011; Nelson *et al.* 2012). This abundance of rare variants influences our ability to infer trait genetic architectures. For example, if rare variants exhibit different population structure than common variants, attempts to control for spatial confounding of genotype and phenotype based on one class of variants will not work for the other class of variants (Mathieson & McVean 2012). Similarly, differences in the distribution of common and rare variants reduce our power to detect the effects of rare functional variants in SNP-based studies with only common genetic variants (Bodmer & Bonilla 2008). This factor might help to explain the problem of missing heritability for quantitative traits and diseases (Cirulli & Goldstein 2010; Eichler *et al.* 2010). Finally, similar problems could stifle attempts to study the genetics of adaptation and speciation, particularly if adaptive traits or barriers to gene flow are highly polygenic and affected by many rare, spatially restricted variants (e.g. Weiss 2008; Rockman 2012). The contribution of rare variants to adaptation or speciation is not well known, but the

© 2014 John Wiley & Sons Ltd

sheer number of rare variants we found in *Lycaeides* suggests that the possible contribution of rare variants to adaptive evolution should not be ignored.

Conclusions

We analysed a large DNA sequence data set to investigate the distribution of genetic and genomic variation within and among *Lycaeides* populations and species. We found that the vast majority of genetic variants in *Lycaeides* were rare or low-frequency variants and that these had different properties than common variants. This has also been shown in humans, and we hypothesize that similar patterns of variation will be found in other organisms. Elucidating the properties of different classes of variants, such as common versus rare variants, is more important now than ever before given our unprecedented ability to obtain genome-wide sequence data from many individuals. With regard to common variants, the spatial and taxonomic distribution of genomic variation in *Lycaeides* butterflies shows evidence of population structure within species and hybridization between species. Hybridization has led to introgression and has been associated with the origin of ecologically distinct admixed lineages. Consequently, patterns of extant variation in this group are multifaceted, with several key axes of genetic variation and ancestry. This complexity, which is not unique to *Lycaeides*, challenges simplistic notions concerning the organization of biological diversity into discrete, easily delineated and hierarchically structured entities.

Acknowledgements

We thank the U.S. National Park (NP) service for providing permits to collect butterflies in Yellowstone NP (YELL-2008-SCI-5682), Grand Teton NP (GRTE-2008-SCI-0024) and Glacier NP (GLAC-2009-SCI-0140). This research was funded by the National Science Foundation (DEB-1050149, DEB-1050355, DEB-1050726, DEB-1050947) and Utah State University. Compute, storage and other resources from the Division of Research Computing in the Office of Research and Graduate Studies at Utah State University are gratefully acknowledged.

References

Anderson EC, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, **160**, 1217–1229.

Balloux F, Lugon-Moulin N (2002) The estimation of population differentiation with microsatellite markers. *Molecular Ecology*, **11**, 155–165.

Barrett RDH, Schluter D (2008) Adaptation from standing genetic variation. *Trends in Ecology and Evolution*, **23**, 38–44.

Barton NH (2001) The role of hybridization in evolution. *Molecular Ecology*, **10**, 551–568.

Barton NH, Slatkin M (1986) A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. *Heredity*, **56**, 409–415.

Bates D, Maechler M, Bolker B, Walker S (2013) *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.0-4.

Begun DJ, Holloway AK, Stevens K, *et al*. (2007) Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biology*, **5**, e310.

Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics*, **40**, 695–701.

Buerkle CA, Rieseberg LH (2008) The rate of genome stabilization in homoploid hybrid species. *Evolution*, **62**, 266–275.

Buerkle CA, Gompert Z, Parchman TL (2011) The n=1 constraint in population genomics. *Molecular Ecology*, **20**, 1575–1581.

de Carvalho D, Ingvarsson PK, Joseph J, *et al*. (2010) Admixture facilitates adaptation from standing variation in the European aspen (*Populus tremula* L.), a widespread forest tree. *Molecular Ecology*, **19**, 1638–1650.

Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, **11**, 415–425.

Clarke R, Rothery P, Raybould A (2002) Confidence limits for regression relationships between distance matrices: Estimating gene flow with distance. *Journal of Agricultural, Biological, and Environmental Statistics*, **7**, 361–372.

Coop G, Pickrell JK, Novembre J, *et al*. (2009) The role of geography in human adaptation. *PLoS Genetics*, **5**, e1000500.

Dasmahapatra KK, Walters JR, Briscoe AD, *et al*. (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, **487**, 94–98.

Eckert AJ, Eckert ML, Hall BD (2010) Effects of historical demography and ecological context on spatial patterns of genetic diversity within foxtail pine (*Pinus balfouriana*; Pinaceae) stands located in the Klamath Mountains, California. *American Journal of Botany*, **97**, 650–659.

Edwards SV, Beerli P (2000) Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, **54**, 1839–1854.

Ehrlich PR, Raven PH (1969) Differentiation of populations. *Science*, **165**, 1228–1232.

Eichler EE, Flint J, Gibson G, *et al*. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, **11**, 446–450.

Ellegren H, Smeds L, Burri R, *et al*. (2012) The genomic landscape of species divergence in *Ficedula flycatchers*. *Nature*, **491**, 756–760.

Endler JA (1977) *Geographic Variation, Speciation, and Clines*. Princeton University Press, Princeton, NewJercy.

Epperson BK (2003) *Geographical Genetics. Monographs in Population Biology*. Princeton University Press, Princeton, NewJercy.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Fitzpatrick BM, Shaffer HB (2007) Hybrid vigor between native and introduced salamanders raises new challenges for conservation. *Proceedings of National Academy of Sciences, USA*, **104**, 15793–15798.

© 2014 John Wiley & Sons Ltd

Fordyce JA, Nice CC (2003) Variation in butterfly egg adhesion: adaptation to local host plant senescence characteristics? *Ecology Letters*, **6**, 23–27.

Fordyce JA, Nice CC, Forister ML, Shapiro AM (2002) The significance of wing pattern diversity in the Lycaenidae: mate discrimination by two recently diverged species. *Journal of Evolutionary Biology*, **15**, 871–879.

Fordyce JA, Gompert Z, Forister ML, Nice CC (2011) A hierarchical Bayesian approach to ecological count data: a flexible tool for ecologists. *PLoS ONE*, **6**, e26785.

Forister ML, Gompert Z, Fordyce JA, Nice CC (2011a) After 60 years, an answer to the question: what is the Karner blue butterfly. *Biology Letters*, **7**, 399–402.

Forister ML, Gompert Z, Nice CC, Forsiter GW, Fordyce JA (2011b) Ant association facilitates the evolution of diet breadth in a Lycaenid butterfly. *Proceedings of the Royal Society B: Biological Sciences*, **278**, 1539–1547.

Forister ML, Scholl CF, Jahner JP, *et al*. (2013) Specificity, rank preference, and the colonization of a non-native host plant by the melissa blue butterfly. *Oecologia*, **172**, 177–188.

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–511.

Girod C, Vitalis R, Leblois R, Fréville H (2011) Inferring population decline and expansion from microsatellite data: a simulation-based evaluation of the msvar method. *Genetics*, **188**, 165–179.

Gompert Z, Buerkle CA (2013) Analyses of genetic ancestry enable key insights for molecular ecology. *Molecular Ecology*, **22**, 5278–5294.

Gompert Z, Fordyce JA, Forister ML, Shapiro AM, Nice CC (2006) Homoploid hybrid speciation in an extreme habitat. *Science*, **314**, 1923–1925.

Gompert Z, Forister ML, Fordyce JA, Nice CC (2008) Widespread mito-nuclear discordance with evidence for introgressive hybridization and selective sweeps in *Lycaeides*. *Molecular Ecology*, **17**, 5231–5244.

Gompert Z, Forister ML, Fordyce JA, Nice CC, Williamson R, Buerkle CA (2010a) Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Molecular Ecology*, **19**, 2455–2473.

Gompert Z, Lucas LK, Fordyce JA, Forister ML, Nice CC (2010b) Secondary contact between *Lycaeides idas* and *L. melissa* in the Rocky Mountains: extensive introgression and a patchy hybrid zone. *Molecular Ecology*, **19**, 3171–3192.

Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, Buerkle CA (2012) Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution*, **66**, 2167–2181.

Gompert Z, Lucas LK, Nice CC, Buerkle CA (2013a) Genome divergence and the genetic architecture of barriers to gene flow between *Lycaeides idas* and *L. melissa*. *Evolution*, **67**, 2498–2514.

Gompert Z, Lucas LK, Nice CC, Fordyce JA, Buerkle CA, Forister ML (2013b) Geographically multifarious phenotypic divergence during speciation. *Ecology and Evolution*, **3**, 595–613.

Grant PR, Grant BR, Markert JA, Keller LF, Petren K (2004) Convergent evolution of Darwin's finches caused by introgressive hybridization and selection. *Evolution*, **58**, 1588–1599.

Gravel S (2012) Population genetics models of local ancestry. *Genetics*, **191**, 607–619.

Gravel S, Henn BM, Gutenkunst RN, *et al*. (2011) Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences, USA*, **108**, 11983–11988.

Guppy C, Shepard J (2001) *Butterflies of British Columbia*. UBC Press, Vancouver, British Columbia.

Harrison RG, Rand DM (1989) Mosaic hybrid zones and the nature of species boundaries. In: *Speciation and Its Consequences* (eds Otte D, Endler J), pp. 110–133. Sinauer Associates, Sunderland, Massachusetts.

Jones FC, Grabherr MG, Chan YF, *et al*. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *PLoS Genetics*, **8**, e1002453.

Lee CR, Mitchell-Olds T (2011) Quantifying effects of environmental and geographical factors on patterns of genetic differentiation. *Molecular Ecology*, **20**, 4631–4642.

Levin DA, Francisco-Ortega JK, Jansen RK (1996) Hybridization and the extinction of rare plant species. *Conservation Biology*, **10**, 10–16.

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

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

Li Y, Vinckenbosch N, Tian G, *et al*. (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature*, **42**, 969–972.

Lucas LK, Fordyce JA, Nice CC (2008) Patterns of genitalic morphology around suture zones in North American Lycaeides (Lepidoptera: Lycaenidae): implications for taxonomy and historical biogeography. *Annals of the Entomological Society of America*, **101**, 172–180.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Mallet J (2008) Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**, 2971–2986.

Mallet J, Beltran M, Neukirchen W, Linares M (2007) Natural hybridization in Heliconiine butterflies: the species boundary as a continuum. *BMC Evolutionary Biology*, **7**, 28.

Mathieson I, McVean G (2012) Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, **44**, 243–246.

Mavárez J, Salazar CA, Bermingham E, Salcedo C, Jiggins CD, Linares M (2006) Speciation by hybridization in Heliconius butterflies. *Nature*, **441**, 868–871.

McDonald DB, Parchman TL, Bower MR, Hubert WA, Rahel FJ (2008) An introduced and a native vertebrate hybridize to form a genetic bridge to a second native species. *Proceedings of National Academy of Sciences, USA*, **105**, 10837–10842.

Molfetti É, Torres Vilaça S, Georges JY, *et al*. (2013) Recent demographic history and present fine-scale structure in the Northwest Atlantic leatherback (*Dermochelys coriacea*) turtle population. *PLoS ONE*, **8**, e58061.

© 2014 John Wiley & Sons Ltd

Nabokov V (1943) The nearctic forms of *Lycaeides* Hub. (Lycaenidae, Lepidoptera). *Psyche*, **50**, 87–99.
Nabokov V (1949) The nearctic members of *Lycaeides* Húbner (Lycaenidae, Lepidoptera). *Bulletin of the Museum of Comparative Zoology*, **101**, 479–541.
Nei M, Maruyama T, Wu C (1983) Models of evolution of reproductive isolation. *Genetics*, **103**, 557–579.
Nelson MR, Wegmann D, Ehm MG, *et al*. (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, **337**, 100–104.
Nice CC, Fordyce JA, Shapiro AM, Ffrench-Constant R (2002) Lack of evidence for reproductive isolation among ecologically specialised lycaenid butterflies. *Ecological Entomology*, **27**, 702–712.
Nice CC, Gompert Z, Fordyce JA, Forister ML, Lucas LK, Buerkle CA (2013) Hybrid speciation and independent evolution in lineages of alpine butterflies. *Evolution*, **67**, 1055–1068.
Nolte AW, Freyhof J, Stemshorn KC, Tautz D (2005) An invasive lineage of sculpins, *Cottus* sp. (Pisces, Teleostei) in the Rhine with new habitat adaptations has originated from hybridization between old phylogeographic groups. *Proceedings of the Royal Society B-Biological Sciences*, **272**, 2379–2387.
Nosil P, Gompert Z, Farkas TE, *et al*. (2012) Genomic consequences of multiple speciation processes in a stick insect. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 5058–5065.
Parchman TL, Gompert Z, Mudge J, Schilkey F, Benkman CW, Buerkle CA (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, **21**, 2991–3005.
Parchman TL, Gompert Z, Braun MJ, *et al*. (2013) The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. *Molecular Ecology*, **22**, 3304–3317.
Patterson N, Moorjani P, Luo Y, *et al*. (2012) Ancient admixture in human history. *Genetics*, **192**, 1065–1093.
Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* , **8**, e1002967.
Plummer M (2003) *JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling*.
Plummer M, Best N, Cowles K, Vines K (2006) CODA: Convergence diagnosis and output analysis for MCMC. *R News*, **5**, 7–11.
Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904–909.
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.
R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing indian population history. *Nature*, **461**, 489–494.
Rieseberg LH, Raymond O, Rosenthal DM, *et al*. (2003) Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, **301**, 1211–1216.
Rockman MV (2012) The QTN program and the alleles that matter for evolution: All that's gold does not glitter. *Evolution*, **66**, 1–17.
Scott J (1986) *The Butterflies of North America: A Natural History and Field Guide*. Stanford University Press, Stanford, California.
Seehausen O (2004) Hybridization and adaptive radiation. *Trends in Ecology and Evolution*, **19**, 198–207.
Seehausen O, Takimoto G, Roy D, Jokela J (2008) Speciation reversal and biodiversity dynamics with hybridization in changing environments. *Molecular Ecology*, **17**, 30–44.
Skotte L, Korneliussen TS, Albrechtsen A (2013) Estimating individual admixture proportions from next generation sequencing data. *Genetics*, **195**, 693–702.
Slatkin M (1985) Rare alleles as indicators of gene flow. *Evolution*, **39**, 53–65.
Slatkin M, Takahata N (1985) The average frequency of private alleles in a partially isolated population. *Theoretical Population Biology*, **28**, 314–331.
Takezaki N, Nei M (1996) Genetic distances and reconstruction of phylogenetic trees from microsatellite dna. *Genetics*, **144**, 389–399.
Taylor EB, Boughman JW, Groenenboom M, Sniatynski M, Schluter D, Gow JL (2006) Speciation in reverse: morphological and genetic evidence of the collapse of a three-spined stickleback (*Gasterosteus aculeatus*) species pair. *Molecular Ecology*, **15**, 343–355.
U.S. Fish and Wildlife Service (2003) *Karner Blue Butterfly (Lycaeides melissa samuelis)*. U.S. Recovery Plan. Tech. rep., Region 3. Fish and Wildlife Service, Fort Snelling, Minnesota.
Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biology*, **4**, e72.
Vonlanthen P, Bittner D, Hudson AG, *et al*. (2012) Eutrophication causes speciation reversal in whitefish adaptive radiations. *Nature*, **482**, 357–362.
Wang IJ, Glor RE, Losos JB (2013) Quantifying the roles of ecology and geography in spatial genetic divergence. *Ecology Letters*, **16**, 175–182.
Weiss KM (2008) Tilting at quixotic trait loci (QTL): an evolutionary perspective on genetic causation. *Genetics*, **179**, 1741–1756.
Whitney KD, Randell RA, Rieseberg LH (2006) Adaptive introgression of herbivore resistance traits in the weedy sunflower Helianthus annuus. *American Naturalist* , **167**, 794–807.
Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences, USA*, **102**, 7882–7887.
Wright S (1943) Isolation by distance. *Genetics*, **28**, 114–138.
Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

All authors designed the research and collected samples. L.K.L., Z.G. and C.C.N. generated the DNA sequence data. Z.G. and C.A.B. contributed new analytical tools. Z.G. analysed the data and wrote the manuscript. All authors revised the manuscript.

© 2014 John Wiley & Sons Ltd

Data accessibility

DNA sequences and alignments (bam format) are archived in NCBI's Sequence Read Archive (SRA) [BioProject ID: PRJNA246037]. Source code for the admixture models, the reference sequence set and the common variant VCF file are available on DRYAD [10.5061/dryad.pq93h].

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 Supplemental Methods and Analyses.

Table S1 Deviance information criterion (DIC) comparing isolation-by-distance (ibd), isolation-by-population (ibp) and full models where $\bar{D}$ is the mean deviance and pD is the effective number of parameters.

Table S2 Three-population test results for admixture in *Lycaeides*.

Table S3 Three-population test results for admixture in *Lycaeides* subgroups.

Fig. S1 Summary of admixture model fit (a) and the cumulative proportion of variance accounted for by individual principal components (b) for common, rare and low-frequency genetic variants.

Fig. S2 Statistical summary of population-genetic structure based on a principal component analysis of low-frequency genetic variants.

Fig. S3 Statistical summary of population-genetic structure based on a principal component analysis of rare genetic variants.

Fig. S4 Statistical summary of population-genetic structure based on a principal component analysis of common genetic variants.

Fig. S5 Admixture proportions based on low-frequency genetic variants.

Fig. S6 Summary of population graph model fits with TREEMIX.

Fig. S7 Population graph inferred by TREEMIX for *Lycaeides* butterflies from 66 localities with alternative roots: melissa-west (a) and anna plus ricei (b).

Fig. S8 Genetic ancestry admixture class estimates (Q) for different taxa and subgroups.

Fig. S9 Distribution of global genetic ancestry estimates (q_1) within and among sample locations for Alpine (a), Warner (b) and Jackson (c) *Lycaeides*.

Fig. S10 Distribution of inter-source population ancestry estimates (Q_{12}) within and among sample locations for Alpine (a), Warner (b) and Jackson (c) *Lycaeides*.

Fig. S11 Genetic ancestry estimates from simulated data.

© 2014 John Wiley & Sons Ltd