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. 2013).
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 Lycaenides butterflies to advance our understanding of the organization and structure of biological diversity. We previously documented patterns of genetic and morphological variation in Lycaenides 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 Lycaenides 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 Lycaenides 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

Lycaenides 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). Lycaenides 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, Lycaenides anna, L. idas and L. melissa (we excluded the endangered Karner blue butterfly, L. samuelis; Forister et al. 2011a, Table 1, Fig. 1a). Lycaenides 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>
<thead>
<tr>
<th>Locality no.</th>
<th>Locality</th>
<th>Taxon</th>
<th>Subgroup</th>
<th>Longitude (° W)</th>
<th>Latitude (° N)</th>
<th>No. of individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Donner Pass, CA</td>
<td>anna</td>
<td>n/a</td>
<td>120.35</td>
<td>39.31</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>Fall Creek, CA</td>
<td>anna</td>
<td>n/a</td>
<td>120.67</td>
<td>39.38</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>Leek Springs, CA</td>
<td>anna</td>
<td>n/a</td>
<td>120.24</td>
<td>38.63</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>Yuba Gap, CA</td>
<td>anna</td>
<td>n/a</td>
<td>120.60</td>
<td>39.32</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>Castle Peak, CA</td>
<td>anna</td>
<td>n/a</td>
<td>120.35</td>
<td>39.37</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>Shovel Creek, CA</td>
<td>ricei</td>
<td>n/a</td>
<td>122.16</td>
<td>41.88</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>Big Lake, OR</td>
<td>ricei</td>
<td>n/a</td>
<td>121.87</td>
<td>44.38</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>Marble Mts., CA</td>
<td>ricei</td>
<td>n/a</td>
<td>122.75</td>
<td>41.83</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>Soda Mt. Road, OR</td>
<td>ricei</td>
<td>n/a</td>
<td>122.48</td>
<td>42.12</td>
<td>19</td>
</tr>
<tr>
<td>10</td>
<td>Cave Lake, CA</td>
<td>ricei</td>
<td>n/a</td>
<td>120.21</td>
<td>41.98</td>
<td>23</td>
</tr>
<tr>
<td>11</td>
<td>Strawberry Mts., OR</td>
<td>idas</td>
<td>n/a</td>
<td>118.64</td>
<td>44.34</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>King’s Hill, MT</td>
<td>idas</td>
<td>n/a</td>
<td>110.70</td>
<td>46.84</td>
<td>18</td>
</tr>
<tr>
<td>13</td>
<td>Soldier Creek, MT</td>
<td>idas</td>
<td>idas-north</td>
<td>114.61</td>
<td>47.21</td>
<td>19</td>
</tr>
<tr>
<td>14</td>
<td>Siyeh Creek, MT</td>
<td>idas</td>
<td>idas-north</td>
<td>113.67</td>
<td>48.70</td>
<td>20</td>
</tr>
<tr>
<td>15</td>
<td>Bunsen Peak, WY</td>
<td>idas</td>
<td>idas-ynp</td>
<td>110.72</td>
<td>44.93</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>Garnet Peak, MT</td>
<td>idas</td>
<td>idas-ynp</td>
<td>111.22</td>
<td>45.43</td>
<td>20</td>
</tr>
<tr>
<td>17</td>
<td>Hayden Valley, WY</td>
<td>idas</td>
<td>idas-ynp</td>
<td>110.49</td>
<td>44.68</td>
<td>22</td>
</tr>
<tr>
<td>18</td>
<td>Tibb’s Butte, WY</td>
<td>idas</td>
<td>n/a</td>
<td>109.45</td>
<td>44.95</td>
<td>20</td>
</tr>
<tr>
<td>19</td>
<td>Animas River, CO</td>
<td>idas</td>
<td>n/a</td>
<td>107.57</td>
<td>37.93</td>
<td>13</td>
</tr>
<tr>
<td>20</td>
<td>Red Mt. Pass, CO</td>
<td>idas</td>
<td>n/a</td>
<td>107.71</td>
<td>37.90</td>
<td>3</td>
</tr>
<tr>
<td>21</td>
<td>Tomboy Road, CO</td>
<td>idas</td>
<td>n/a</td>
<td>107.77</td>
<td>37.94</td>
<td>24</td>
</tr>
<tr>
<td>22</td>
<td>Bishop, CA</td>
<td>melissa</td>
<td>melissa-west</td>
<td>118.28</td>
<td>37.17</td>
<td>20</td>
</tr>
<tr>
<td>23</td>
<td>Gardnerville, NV</td>
<td>melissa</td>
<td>melissa-west</td>
<td>119.78</td>
<td>38.81</td>
<td>17</td>
</tr>
<tr>
<td>24</td>
<td>Red Earth Way, NV</td>
<td>melissa</td>
<td>melissa-west</td>
<td>118.84</td>
<td>38.98</td>
<td>20</td>
</tr>
<tr>
<td>25</td>
<td>Silver Lake, NV</td>
<td>melissa</td>
<td>melissa-west</td>
<td>119.93</td>
<td>39.65</td>
<td>18</td>
</tr>
<tr>
<td>26</td>
<td>Verdi, NV</td>
<td>melissa</td>
<td>melissa-west</td>
<td>120.00</td>
<td>39.51</td>
<td>20</td>
</tr>
<tr>
<td>27</td>
<td>Washoe Lake, NV</td>
<td>melissa</td>
<td>melissa-west</td>
<td>118.82</td>
<td>38.65</td>
<td>8</td>
</tr>
<tr>
<td>28</td>
<td>De Beque, CO</td>
<td>melissa</td>
<td>melissa-east</td>
<td>108.21</td>
<td>39.32</td>
<td>20</td>
</tr>
<tr>
<td>29</td>
<td>Deeth-Charleston, NV</td>
<td>melissa</td>
<td>melissa-east</td>
<td>115.38</td>
<td>41.30</td>
<td>20</td>
</tr>
<tr>
<td>30</td>
<td>Goose Lake, CA</td>
<td>melissa</td>
<td>melissa-east</td>
<td>120.29</td>
<td>41.99</td>
<td>20</td>
</tr>
<tr>
<td>31</td>
<td>Lamoille Canyon, NV</td>
<td>melissa</td>
<td>melissa-east</td>
<td>115.47</td>
<td>40.68</td>
<td>19</td>
</tr>
<tr>
<td>32</td>
<td>Ophir City, NV</td>
<td>melissa</td>
<td>melissa-east</td>
<td>117.27</td>
<td>38.94</td>
<td>19</td>
</tr>
<tr>
<td>33</td>
<td>Star Creek Canyon, NV</td>
<td>melissa</td>
<td>melissa-east</td>
<td>118.12</td>
<td>40.55</td>
<td>16</td>
</tr>
<tr>
<td>34</td>
<td>Surprise Valley, CA</td>
<td>melissa</td>
<td>melissa-east</td>
<td>120.10</td>
<td>41.28</td>
<td>20</td>
</tr>
<tr>
<td>35</td>
<td>Montague, CA</td>
<td>melissa</td>
<td>melissa-east</td>
<td>122.38</td>
<td>41.77</td>
<td>19</td>
</tr>
<tr>
<td>36</td>
<td>Cokeville, WY</td>
<td>melissa</td>
<td>melissa-east</td>
<td>110.94</td>
<td>42.01</td>
<td>10</td>
</tr>
<tr>
<td>37</td>
<td>Montrose, CO</td>
<td>melissa</td>
<td>melissa-east</td>
<td>107.82</td>
<td>38.37</td>
<td>20</td>
</tr>
<tr>
<td>38</td>
<td>Victor, ID</td>
<td>melissa</td>
<td>melissa-east</td>
<td>111.11</td>
<td>43.66</td>
<td>20</td>
</tr>
<tr>
<td>39</td>
<td>Cody, WY</td>
<td>melissa</td>
<td>melissa-rockies</td>
<td>108.98</td>
<td>44.51</td>
<td>23</td>
</tr>
<tr>
<td>40</td>
<td>Lander, WY</td>
<td>melissa</td>
<td>melissa-rockies</td>
<td>108.36</td>
<td>42.65</td>
<td>24</td>
</tr>
<tr>
<td>41</td>
<td>Yellow Pine, WY</td>
<td>melissa</td>
<td>melissa-rockies</td>
<td>105.40</td>
<td>41.25</td>
<td>19</td>
</tr>
<tr>
<td>42</td>
<td>Albion Meadows, UT</td>
<td>melissa</td>
<td>melissa-rockies</td>
<td>111.62</td>
<td>40.59</td>
<td>46</td>
</tr>
<tr>
<td>43</td>
<td>Mt. Rose, NV</td>
<td>alpine</td>
<td>mt-rocke</td>
<td>119.93</td>
<td>39.32</td>
<td>52</td>
</tr>
<tr>
<td>44</td>
<td>Carson Pass, CA</td>
<td>alpine</td>
<td>carson</td>
<td>120.02</td>
<td>38.71</td>
<td>50</td>
</tr>
<tr>
<td>45</td>
<td>Corev Peak, NV</td>
<td>alpine</td>
<td>sierra-nevada</td>
<td>118.77</td>
<td>38.45</td>
<td>8</td>
</tr>
<tr>
<td>46</td>
<td>Lake Emma, CA</td>
<td>alpine</td>
<td>sierra-nevada</td>
<td>119.48</td>
<td>38.28</td>
<td>32</td>
</tr>
<tr>
<td>47</td>
<td>Sonora Pass, CA</td>
<td>alpine</td>
<td>sierra-nevada</td>
<td>119.63</td>
<td>38.33</td>
<td>44</td>
</tr>
<tr>
<td>48</td>
<td>Sweetwater Mts., CA</td>
<td>alpine</td>
<td>sierra-nevada</td>
<td>119.33</td>
<td>38.45</td>
<td>23</td>
</tr>
<tr>
<td>49</td>
<td>Tioga Crest, CA</td>
<td>alpine</td>
<td>sierra-nevada</td>
<td>119.26</td>
<td>37.97</td>
<td>38</td>
</tr>
<tr>
<td>50</td>
<td>South Fork, CA</td>
<td>alpine</td>
<td>white</td>
<td>118.57</td>
<td>37.21</td>
<td>14</td>
</tr>
<tr>
<td>51</td>
<td>Reed Flat, CA</td>
<td>alpine</td>
<td>white</td>
<td>118.18</td>
<td>37.38</td>
<td>8</td>
</tr>
<tr>
<td>52</td>
<td>County Line Hill, CA</td>
<td>alpine</td>
<td>white</td>
<td>118.19</td>
<td>37.46</td>
<td>40</td>
</tr>
<tr>
<td>53</td>
<td>Buck Mt., CA</td>
<td>warner</td>
<td>n/a</td>
<td>120.29</td>
<td>41.69</td>
<td>44</td>
</tr>
<tr>
<td>54</td>
<td>Eagle Peak, CA</td>
<td>warner</td>
<td>n/a</td>
<td>120.22</td>
<td>41.26</td>
<td>40</td>
</tr>
</tbody>
</table>
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-ypn’) 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.
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 SEQUAN 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.

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).
(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 Lycaenidae 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
ADMIXTURE AND DIVERSITY IN A SPECIES COMPLEX 4561

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 *Lycaedes* populations is consistent with a bifurcating tree. We first used the TREPIM method proposed by Pickrell & Pritchard (2012) to infer a population graph representing the hypothesized evolutionary history of the sampled *Lycaedes* populations. Population graphs allow both population splits and admixture or migration events. We constructed *Lycaedides* 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*2-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
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 and preceded by 5000 iteration burn-ins, for each set of 15 000 iteration chains, sampled every fifth iteration (e.g. PCA and admixture proportion estimates; see Supporting information).

**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
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
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 Q11), (ii) both gene copies were inherited from an alternative source population (nonzero Q22) and (iii) each gene copy was inherited from a different source population (nonzero Q12; Fig. S8, Supporting information). We found less evidence of inter-source population ancestry (i.e. nonzero Q12) 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[Q12] = 2q1[1−q1]; 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 (q1; 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 (q1) 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 (Q12) 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).

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).
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.
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

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.

© 2014 John Wiley & Sons Ltd
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’
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,
Ancient admixture

Recent admixture

Repeated admixture

Structured

Evo. independence

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
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-5690), 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


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 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 (q1) within and among sample locations for Alpine (a), Warner (b) and Jackson (c) Lycaeides.

Fig. S10 Distribution of inter-source population ancestry estimates (Q12) within and among sample locations for Alpine (a), Warner (b) and Jackson (c) Lycaeides.

Fig. S11 Genetic ancestry estimates from simulated data.